Fluid Limits for Bandwidth-Sharing Networks with Rate Constraints

Maria Frolkova, Josh Reed and Bert Zwart*

February 12, 2012

Abstract

Bandwidth-sharing networks as introduced by Massoulié & Roberts (1998) model the dynamic interaction among an evolving population of elastic flows competing for several links. With policies based on optimization procedures, such models are of interest both from a Queueing Theory and Operations Research perspective.

In the present paper, we focus on bandwidth-sharing networks with capacities and arrival rates of a large order of magnitude compared to transfer rates of individual flows. This regime is standard in practice. In particular, we extend previous work by Reed & Zwart (2010) on fluid approximations for such networks: we allow interarrival times, flow sizes and patience times to be generally distributed, rather than exponentially distributed. We also develop polynomial-time computable fixed-point approximations for stationary distributions of bandwidth-sharing networks, and suggest new techniques for deriving these types of results.

Keywords: bandwidth-sharing, rate constraints, impatience, large capacity scaling, fluid limits, fixed-point approximations.

MSC2010: Primary 60K25, 60K30, 60F17, 60G57; Secondary 90B15, 90B22.

1 Introduction

Bandwidth-sharing policies as introduced by Massoulié & Roberts [25, 21] dynamically distribute network resources among a changing population of users. Processor sharing is an example of such a policy and assumes a single resource. Bandwidth-sharing networks are of great research and practical interest. Along with the basic application in telecommunications media, e.g. Internet congestion control, they also have recently been suggested as a tool in analyzing problems in road traffic [19].

The main issues in bandwidth-sharing related research are stability conditions and performance evaluation. A variety of results regarding the first topic may be found in De Veciana *et al.* [27, 28], Bonald & Massoulié [5], Mo & Walrand [22], Massoulié [20], Bramson [7], Gromoll & Williams [12], and Chiang *et al.* [8]. As for the second topic, for special combinations of network topologies and bandwidth-sharing policies, the network stationary distribution may be shown to be of a product form insensitive to the flow size distribution, see Bonald *et al.* [6]. However, in general, approximation methods must be used, which is the subject matter of the present paper. Fundamental papers on fluid limit approximations for bandwidth sharing-networks are

*MF is with CWI, P.O. Box 94079, 1098 XG Amsterdam, The Netherlands. E-mail: M.Frolkova@cwi.nl. JR is with NYU Stern School of Business, 44 West 4th St., Suite 8-79, New York, NY 10012, the USA. E-mail: jreed@stern.nyu.edu. BZ is with CWI. E-mail: Bert.Zwart@cwi.nl. BZ is also affiliated with EURANDOM, VU University Amsterdam, and Georgia Institute of Technology. The research of MF and BZ in supported by an NWO VIDI grant.

Kelly & Williams [18] and Gromoll & Williams [13], some more results on fluid and diffusion approximations are to be found in Borst *et al.* [9, 4], Kang *et al.* [16] and Ye & Yao [29, 30]. The latter works ignore the impact of individual peak rate limitations though, as has been pointed out by Roberts [24].

To the best of our knowledge, Ayesta & Mandjes [2] were the first to deal with fluid and diffusion approximations of bandwidth-sharing networks with peak rate limitations. They consider two specific settings first without rate constraints, and then they truncate the capacity constraints at the rate maxima. Reed & Zwart [23] develop a different approach in the context of general bandwidth-sharing networks. They incorporate the rate constraints into the network utility maximization procedure that defines bandwidth allocations. Thus, users operating below the maximal rate are allowed to take up the bandwidth that is not used by other rate constrained users, and bandwidth allocations are Pareto optimal. Another interesting feature of this work is the scaling regime. In contrast to the papers mentioned above, which mostly focus on the large-time properties of networks with fixed-order parameters, Reed & Zwart view networks on a fixed-time scale letting arrival rates and capacities grow large. This large capacity scaling reflects the fact that overall network capacity and individual user rate constraints may be of different orders of magnitude. For example, it is common that Internet providers set download speed limitations for individual users which are typically measured in megabits per second, while network capacities are measured in gigabits or terabits per second.

The present paper builds upon [23] by relaxing its stochastic assumptions: we assume general distribution for interarrival times and general joint distribution for the size and patience time of a flow (in particular, the flow size and patience time are allowed to be dependent), while [23] assumes a Markovian setting with independent arrivals, flow sizes and patience times. We study the behavior of bandwidth-sharing networks in terms of measure-valued processes that are called state descriptors and that keep track of residual flow sizes and residual patience times. The first main result of the paper is a fluid limit theorem (it generalizes the fluid limit result of [23] to non-Markovian stochastic assumptions). We propose a fluid model, or a formal deterministic approximation of the stochastic bandwidth-sharing model, and show that the scaled state descriptors are tight with all weak limit points a.s. solving the fluid model equation. We provide a sufficient condition for the fluid model to have a unique solution, which converts tightness of the scaled state descriptors into convergence to this fluid model solution. In the sense of techniques used in the proofs, this part of the paper is closely related to previous work on bandwidth-sharing [13], processor-sharing with impatience [11], and bandwidth-sharing in overload [4, 9]. The rate constraints play a crucial role in adopting these techniques. For example, the proof of convergence to fluid model solutions in [11] requires an additional assumption of overload to eliminate problems at zero. However, in our case, due to the rate constraints, the network never empties, and the load conditions become irrelevant.

Our second main result, which is a new type of result for bandwidth-sharing networks, is convergence of the scaled network stationary distribution to the fixed point of the fluid model, provided the fixed point is unique. There is a similar result by Kang & Ramanan [17] for a call center model, but the techniques of [17] are different than ours. Applying the approach of Borst *et al.* [4], we prove that in many cases the fixed point can be found by solving an optimization problem with a strictly concave objection function and a polyhedral constraint set, and thus is unique and computable in polynomial time. We also construct an example with multiple fixed points, which is a feature that is distinctive from earlier cited works. Besides proving new results for the particular model of bandwidth-sharing, we also suggest new ideas and believe that they can be adjusted to other models, too. In particular, we derive equations for asymptotic bounds for fluid model solutions (see Theorem 3) that can be solved for a wide class of networks, and then asymptotic stability of the fixed point can be shown. Another interesting idea is that, in the stationary regime, the properties of a network depend on newly arriving flows only, since all initial flows are gone after some point (see Lemma 2). Throughout this part of the paper,

we assume Poisson arrivals, since that guarantees existence of a unique stationary distribution. Poisson arrivals also imply $M/G/\infty$ bounds that are exploited heavily in the proofs.

The structure of the paper is as follows. Section 2 describes the stochastic bandwidth-sharing model, and Section 3 introduces its deterministic analogue, the fluid model. Also Section 3 states sufficient conditions for a fluid model solution to be unique, and for a fixed fluid model solution to be unique and asymptotically stable. Sections 4 and 5 discuss convergence of the scaled state descriptor and its stationary distribution to the fluid model and its fixed point, respectively. Sections 6, 7 and 8 contain the proofs of the statements from Sections 3, 4 and 5. The Appendix proves auxiliary results. In the remainder of this section, we list the notation we use throughout the paper.

Notation In order to introduce the notation, we use the signs $=:$ and $:=$.

The standard sets are denoted as follows: the reals $\mathbb{R} = (-\infty, \infty)$, the non-negative reals $\mathbb{R}_+ = [0, \infty)$, the positive reals $(0, \infty)$, the non-negative integers $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$, and the natural numbers $\mathbb{N} = \{1, 2, \dots\}$.

The signs $\wedge$ and $\vee$ stand for minimum and maximum respectively. For $x \in \mathbb{R}$, $x^+ := x \vee 0$.

The signs $\underline{\lim}$ and $\overline{\lim}$ denote the lower and upper limits of a sequence of numbers.

The coordinates of a vector from a set S^I are denoted by the same symbol as the vector with lower indices $1, \dots, I$ added. If a vector has a superscript, tilde-sign, or overlining, they remain in its coordinates. For example $\overline{x}^0 \in S^d$, $\overline{x}^0 = (\overline{x}^0_1, \dots, \overline{x}^0_I)$. The space $\mathbb{R}^I$ is endowed with the supremum norm $\|x\| := \max_{1 \le i \le I} |x_i|$. Vector inequalities hold coordinate-wise. The coordinate-wise product of vectors of the same dimensionality I is $x * y := (x_1 y_1, \dots, x_I y_I)$.

The signs $\Rightarrow$, $\stackrel{\mathrm{d}}{=}$ and $\le_{\mathrm{st}}$ stand for convergence in distribution, equality in distribution and stochastic order respectively. Recall that, for real-valued r.v.'s X and X', $X \le_{\mathrm{st}} X'$ if $\mathbb{P}\{X > x\} \le \mathbb{P}\{X' > x\}$ for all $x \in \mathbb{R}$. The notation $\Pi(\lambda)$, $\lambda \in (0, \infty)$, stands for the Poisson distribution with parameter λ.

For metric spaces S and S', denote by $\mathbf{C}_{S \to S'}$ the space of continuous functions $f \colon S \to S'$. By $\mathbf{D}_{\mathbb{R}_+ \to S}$ denote the space of functions $f \colon \mathbb{R}_+ \to S$ that are right-continuous with left limits, and endow this space with the Skorokhod J_1-topology.

The superscript -1 is only used to denote the inverse of a function.

For a measure ξ on $\mathbb{R}^2_+$ and a ξ-integrable function $f \colon \mathbb{R}^2_+ \to \mathbb{R}$, define $\langle f, \xi \rangle := \int_{\mathbb{R}^2_+} f d\xi$. If $\xi = (\xi_1, \dots, \xi_I)$ is a vector of such measures, $\langle f, \xi \rangle := (\langle f, \xi_1 \rangle, \dots, \langle f, \xi_I \rangle)$. Let $\mathbf{M}$ be the space of finite non-negative Borel measures on $\mathbb{R}^2_+$ endowed with the weak topology: $\xi^k \stackrel{\mathrm{w}}{\to} \xi$ in $\mathbf{M}$ as $k \to \infty$ if and only if $\langle f, \xi^k \rangle \to \langle f, \xi \rangle$ for all continuous bounded function $f \colon \mathbb{R}^2_+ \to \mathbb{R}$. Weak convergence of elements of $\mathbf{M}$ is equivalent to convergence in the Prokhorov metric: for $\xi, \varphi \in \mathbf{M}$, define

$$d(\xi, \varphi) := \inf\{\varepsilon \colon \xi(B) \le \varphi(B^\varepsilon) + \varepsilon \text{ and } \varphi(B) \le \xi(B^\varepsilon) + \varepsilon \text{ for all non-empty closed } B \subseteq \mathbb{R}^2_+\},$$

where $B^\varepsilon = \{x \in \mathbb{R}^2_+ \colon \inf_{y \in B} \|x - y\| < \varepsilon\}$.

For $\xi, \varphi \in \mathbf{M}^I$, define

$$d_I(\xi, \varphi) := \max_{1 \le i \le I} d(\xi_i, \varphi_i).$$

Equipped with the metric $d_I(\cdot, \cdot)$, the space $\mathbf{M}^I$ is separable and complete.

2 Stochastic model

This section contains a detailed description of the model under consideration. In particular, it specifies the structure of the network, the policy it operates under and the stochastic dynamical assumptions. Also, a stochastic process is introduced that keeps track of the state of the network, see the state descriptor paragraph.

Network structure Consider a network that consists of a finite number of links labeled by $j = 1, \dots, J$. Traffic offered to the network is represented by elastic flows coming from a finite number of classes labeled by $i = 1, \dots, I$. All class i flows are transferred through a certain subset of links, we call it *route* i. Transfer of a flow starts immediately upon its arrival and is continuous with all links on the route of the flow being traversed simultaneously. Let A be the $J \times I$ incidence matrix, where $A_{ji} = 1$ if route i contains link j and $A_{ji} = 0$ otherwise.

Suppose that at a particular time t the population of the network is $z \in \mathbb{Z}_+^I$, where z_i stands for the number of flows on route i. All flows on route i are transferred at the same rate $\lambda_i(z)$ that is at most $m_i \in (0, \infty)$. If $z_i = 0$, put $\lambda_i(z) := 0$. We refer to $\Lambda_i(z) := \lambda_i(z) z_i$ as the *bandwidth allocated to route* i. The sum of the bandwidths allocated to the routes that contain link j is the *bandwidth allocated through link* j and is at most $C_j \in (0, \infty)$. We call C_j the *capacity of link* j. Hence, the vectors $\lambda(z) = (\lambda_1(z), \dots, \lambda_I(z))$ and $\Lambda(z) = (\Lambda_1(z), \dots, \Lambda_I(z))$ must satisfy

$$A(\lambda(z) \cdot z) = A\Lambda(z) \le C, \quad \lambda(z) \le m, \quad \Lambda(z) \le m * z,$$

where $C = (C_1, \dots, C_J)$ and $m = (m_1, \dots, m_I)$ are the vectors of link capacities and rate constraints.

Bandwidth-sharing policy At each point in time, the link capacities should be distributed among the routes in such a way that the network utility is maximized. Namely, to each flow on route i we assign a utility $\mathcal{U}_i(\cdot)$ that is a function of the rate allocated to that flow. Assume that the functions $\mathcal{U}_i(\cdot)$ are strictly increasing and concave in $\mathbb{R}_+$, and twice differentiable in $(0, \infty)$ with $\lim_{x \downarrow 0} \mathcal{U}_i'(x) = \infty$. We also allow $\lim_{x \downarrow 0} \mathcal{U}_i(x) = -\infty$ as, for example, in the case of a logarithmic function. Then, for $z \in \mathbb{R}_+^I$, the vector $\lambda(z)$ of rates is the unique optimal solution to

$$\text{maximize} \quad \sum\nolimits_{i=1}^{I} z_i \mathcal{U}_i(\lambda_i) \quad \text{subject to} \quad A(\lambda * z) \le C, \quad \lambda \le m, \tag{1}$$

where, by convention, $0 \times (-\infty) := 0$. Although the population vector has integer-valued coordinates, we assume that $\lambda(z)$ and $\Lambda(z) := \lambda(z) * z$ are defined via (1) in the entire orthant $\mathbb{R}_+^I$ to accommodate fluid analogues of the population process later.

The utility maximization procedure (1) implies that $\lambda_i(z) = \Lambda_i(z) = 0$ if $z_i = 0$. The assumption $\lim_{x \downarrow 0} \mathcal{U}_i'(x) = \infty$ guarantees non-idling, that is $\lambda_i(z), \Lambda_i(z) > 0$ if $z_i > 0$. Reed & Zwart [23] proved that the functions $\lambda(\cdot)$ and $\Lambda(\cdot)$ are differentiable in any direction and, in particular, locally Lipschitz continuous in the interior of $\mathbb{R}_+^I$. We also show continuity of $\Lambda(\cdot)$ on the boundary of $\mathbb{R}_+^I$ (see the Appendix).

Lemma 1. *The bandwidth allocation function* $\Lambda(\cdot)$ *is continuous in* $\mathbb{R}_+^I$.

Stochastic assumptions All stochastic primitives introduced in this paragraph are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with expectation operator $\mathbb{E}$.

Suppose at time zero there is an a.s. finite number of flows in the network, we call them *initial flows*. A random vector $Z^0 \in \mathbb{R}_+^I$ represents the initial population, and Z_i^0 is the number of initial flows on route i. New flows arrive to the network according to a stochastic process $E(\cdot) = (E_1(\cdot), \dots, E_I(\cdot))$ with sample paths in the Skorokhod space $\mathbf{D}_{\mathbb{R}_+ \to \mathbb{R}_+^I}$. The coordinates of the

arrival process are independent counting processes. Recall that a *counting process* is a non-decreasing non-negative integer-valued process starting from zero. For $t \geq 0$, $E_i(t)$ represents the number of flows that have arrived to route i during the time interval $(0, t]$. The kth such arrival occurs at time $U_{ik} = \inf\{t \geq 0\colon E_i(t) \geq k\}$, it is called *flow k on route i*, $k \in \mathbb{N}$. Simultaneous arrivals are allowed.

Flows abandon the network due to transfer completions or because they run out of patience, depending on what happens earlier for each particular flow. Flow sizes and patience times are drawn from sequences $\{(B^0_{il}, D^0_{il})\}_{l\in\mathbb{N}}$, $\{(B_{ik}, D_{ik})\}_{k\in\mathbb{N}}$, $i = 1, \ldots, I$, of $(0,\infty)^2$-valued r.v.'s. For $l = 1, \ldots, Z^0_i$, B^0_{il} and D^0_{il} represent the residual size and residual patience time at time zero of initial flow l on route i. For $k \in \mathbb{N}$, B_{ik} and D_{ik} represent the initial size and initial patience time of flow k on route i, where "initial" means as upon arrival at time U_{ik}. Let (B_{ik}, D_{ik}), $k \in \mathbb{N}$, be i.i.d. copies of a r.v. (B_i, D_i) with distribution law θ_i and a finite mean value $(1/\mu_i, 1/\nu_i)$. Also introduce the notations $\rho_i := \eta_i/\mu_i$ and $\sigma_i := \eta_i/\nu_i$. Assume that the sequences $\{(B_{ik}, D_{ik})\}_{k\in\mathbb{N}}$ are independent and do not depend on the arrival process $E(\cdot)$. For the moment, we do not make any specific assumptions about the sequences $\{(B^0_{il}, D^0_{il})\}_{l\in\mathbb{N}}$.

State descriptor We denote the *population process* by $Z(\cdot) = (Z_1(\cdot), \ldots, Z_I(\cdot))$, where $Z_i(t)$ is the number of flows on route i at time t. As can be seen from what follows, $Z(\cdot)$ is a random element of the Skorokhod space $\mathbf{D}_{\mathbb{R}_+\to\mathbb{R}^I_+}$.

For $i = 1, \ldots, I$, introduce operators $S_i\colon \mathbf{D}_{\mathbb{R}_+\to\mathbb{R}^I_+} \times \mathbb{R}^2_+ \to \mathbf{C}_{\mathbb{R}^2_+\to\mathbb{R}_+}$ defined by

$$S_i(z, s, t) := \int_s^t \lambda_i(z(u))du,$$

For $t \geq s \geq 0$, $S_i(Z, s, t)$ is the *cumulative bandwidth allocated per flow on route i during time interval* $[s, t]$. The *residual size* and *residual lead time at time t* of initial flow $l = 1, \ldots, Z^0_i$ on route i are given by

$$B^0_{il}(t) := (B^0_{il} - S_i(Z, 0, t))^+ \quad \text{and} \quad D^0_{il}(t) := (D^0_{il} - t)^+,$$

and those of flow $k = 1, \ldots, E_i(t)$ on route i by

$$B_{ik}(t) := (B_{ik} - S_i(Z, U_{ik}, t))^+ \quad \text{and} \quad D_{ik}(t) := (D_{ik} - (t - U_{ik}))^+.$$

The state of the network at any time t is defined by the residual sizes and residual patience times of the flows present in the network. With each flow, we associate a dot in $\mathbb{R}^2_+$, whose coordinates are the residual size and residual patience time of the flow (see Fig. 1). As a flow is getting transferred, the corresponding dot moves toward the axis: to the left at the transfer rate (which is $\lambda_i(Z(t))$ for a flow on route i) and downward at the constant rate of 1. As a dot hits the vertical axis, the corresponding flow leaves due to completion of its transfer. As a dot hits the horizontal axis, the corresponding flow leaves due to impatience. We combine these moving dots into the stochastic process $\mathcal{Z}(\cdot) \in \mathbf{D}_{\mathbb{R}_+\to\mathbf{M}^I}$ with

$$\mathcal{Z}_i(t) := \sum\nolimits_{l=1}^{Z^0_i} \delta^+_{(B^0_{il}(t), D^0_{il}(t))} + \sum\nolimits_{k=1}^{E_i(t)} \delta^+_{(B_{ik}(t), D_{ik}(t))}, \tag{2}$$

where, for $x \in \mathbb{R}^2_+$, $\delta^+_x \in \mathbf{M}$ is the Dirac measure at x if $x_1 \wedge x_2 > 0$ and zero measure otherwise (i.e. assigns a zero mass to any Borel subset of $\mathbb{R}^2_+$). That is, $\mathcal{Z}_i(t)$ is a counting measure on $\mathbb{R}^2_+$ that puts a unit mass to each of the dots representing class i flows except those on the axis. The process $\mathcal{Z}(\cdot)$ given by (2) we call the *state descriptor*. Note that the total mass of the state descriptor coincides with the network population, $\langle 1, \mathcal{Z}(\cdot)\rangle = Z(\cdot)$.

When proving the results of the paper, we decompose the state descriptors into two parts keeping track of initial and newly arriving flows, respectively. That is,

$$\mathcal{Z}(\cdot) = \mathcal{Z}^{\text{init}}(\cdot) + \mathcal{Z}^{\text{new}}(\cdot),$$

where

$$\mathcal{Z}_i^{\rm init}(t) := \sum_{l=1}^{Z_i(0)} \delta^+_{(B^0_{il}(t),D^0_{il}(t))} \quad \text{and} \quad \mathcal{Z}_i^{\rm new}(t) := \sum_{k=1}^{E_i(t)} \delta^+_{(B_{ik}(t),D_{ik}(t))}.$$

We also define the corresponding total mass processes

$$Z^{\rm init}(\cdot) := \langle 1, \mathcal{Z}^{\rm init}(\cdot)\rangle \quad \text{and} \quad Z^{\rm new}(\cdot) := \langle 1, \mathcal{Z}^{\rm new}(\cdot)\rangle.$$

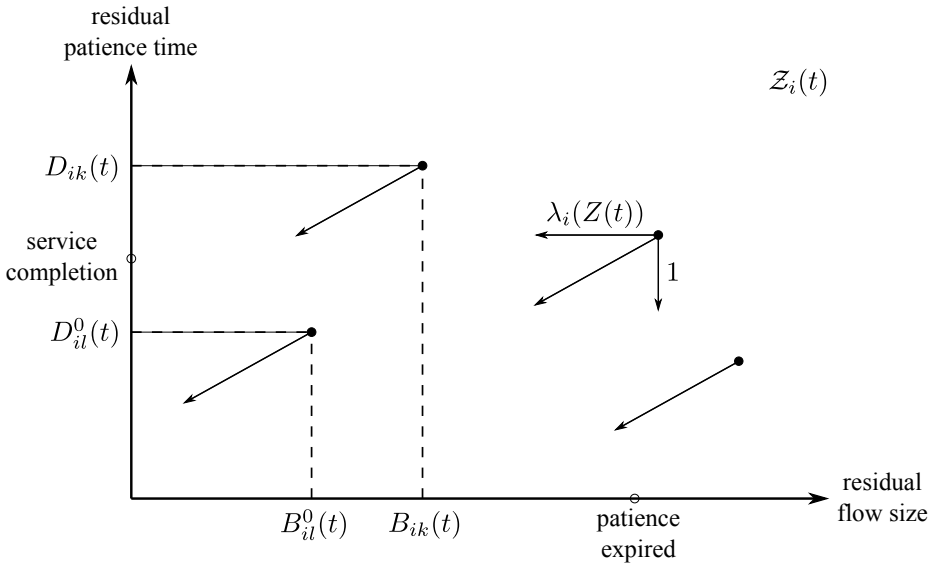

Figure 1: The i-th coordinate $\mathcal{Z}_i(\cdot)$ of the state descriptor puts a unit mass to the dots representing class i flows except those on the axis.

3 Fluid model

In this section we define and investigate a fluid model that is a deterministic analogue of the stochastic model described in the previous section. Later on the fluid model will be shown to arise as the limit of the stochastic model under a proper scaling. This convergence implies, in particular, existence of the fluid model.

To define the fluid model we need data $(\eta, \theta, \zeta^0) \in (0,\infty)^I \times \mathbf{M}^I \times \mathbf{M}^I$. The coordinates of η play the role of arrival rates. Recall that θ_i is the joint distribution of the generic size B_i and patience time D_i of a newly arrived flow on route i. Finally, ζ^0 characterizes the initial state of the network. Also put $z^0 := \langle 1, \zeta^0\rangle$ and, for all i, take a r.v. (B^0_i, D^0_i) that is degenerate at $(0,0)$ if $z^0_i = 0$ and has distribution ζ^0_i/z^0_i otherwise. Then z^0 represents the initial population, and (B^0_i, D^0_i) the generic size and patience time of an initial flow on route i.

Denote by $\mathcal{C}$ the collection of corner sets,

$$\mathcal{C} := \{[x,\infty)\times[y,\infty)\colon (x,y)\in\mathbb{R}^2_+\}.$$

Definition 1. A pair $(\zeta, z) \in \mathbf{C}_{\mathbb{R}_+\to\mathbf{M}^I} \times \mathbf{C}_{\mathbb{R}_+\to\mathbb{R}^I_+}$ is called a *fluid model solution (FMS) for the data* (η,θ,ζ^0) if $z(\cdot) = \langle 1, \zeta(\cdot)\rangle$ and, for all i, $t \ge 0$ and $A \in \mathcal{C}$,

$$\begin{aligned}\zeta_i(t)(A) =& z^0_i \mathbb{P}\{(B^0_i, D^0_i) \in A + (S_i(z,0,t),t)\}\\ &+ \eta_i \int_0^t \mathbb{P}\{(B_i,D_i)\in A + (S_i(z,s,t), t-s)\}ds. \end{aligned} \tag{3}$$

In particular, for all i and $t \geq 0$,

$$z_i(t) = \zeta_i(t)(\mathbb{R}^2_+) = z_i^0 \mathbb{P}\{B_i^0 \geq S_i(z,0,t), D_i^0 \geq t\} + \eta_i \int_0^t \mathbb{P}\{B_i \geq S_i(z,s,t), D_i \geq t-s\} ds. \tag{4}$$

The function $\zeta(\cdot)$ is called a *measure-valued fluid model solution (MVFMS)* and the function $z(\cdot)$ a *numeric fluid model solution (NFMS)*

Equations (3) and (4) have appealing physical interpretations. For example, (4) simply means that a flow is still in the network at time t if its size and patience exceed, respectively, the amount of service it has received and the time that has passed since its arrival up to time t.

Remark 1. By Dynkin's π-λ theorem (see [11, Section 2.3]), FMS's satisfy (3) with any Borel set $A \subseteq \mathbb{R}^2_+$.

Remark 2. FMS's are invariant with respect to time shifts in the sense that, if $(\zeta, z)(\cdot)$ is an FMS, then, for any $\delta > 0$, $(\zeta^\delta, z^\delta)(\cdot) := (\zeta, z)(\cdot + \delta)$ is an FMS for the data $(\eta, \theta, \zeta(\delta))$. That is, for all i, $t \geq \delta$ and Borel sets $A \subseteq \mathbb{R}^2_+$,

$$\zeta_i(t)(A) = \zeta_i(\delta)(A + (S_i(z,\delta,t), t-\delta)) + \eta_i \int_\delta^t \mathbb{P}\{(B_i, D_i) \in A + (S_i(z,s,t), t-s)\} ds, \tag{5a}$$

$$z_i(t) = \zeta_i(\delta)([S_i(z,\delta,t), \infty) \times [t-\delta, \infty)) + \eta_i \int_\delta^t \mathbb{P}\{B_i \geq S_i(z,s,t), D_i \geq t-s\} ds. \tag{5b}$$

Remark 3. The measure-valued and numeric components of an FMS uniquely define each other. In particular, uniqueness of an NFMS implies uniqueness of an MVFMS, and the other way around.

Further we discuss sufficient conditions for an FMS to be unique and for an invariant (i.e. constant) FMS to be unique and asymptotically stable. To prove the stability result, we derive relations for asymptotic bounds for FMS's, which seems to be a novel approach since we have not seen analogous results in the related literature. We also give an example of multiple invariant FMS's.

Uniqueness of an FMS The proof of the following theorem follows along the lines of the proofs of similar results [4, Proposition 4.2] and [11, Theorem 3.5], see Section 6.

Theorem 1. *Suppose that either* (i) $z_i^0 = 0$ *for all* i, *or* (ii) $z^0 \in (0,\infty)^I$ *and the first projection of* ζ^0 *is Lipschitz continuous, i.e. there exists a constant* $L \in (0,\infty)$ *such that for all* i, $x < x'$ *and* y,

$$\zeta_i^0([x, x'] \times [y, \infty)) \leq L(x' - x).$$

Then an FMS for the data (η, θ, ζ^0) *is unique.*

Uniqueness of an invariant FMS Let (ζ, z) be an invariant FMS. By Lemma 3 in Section 6, all of the coordinates of z are positive, and the fluid model equations (3) and (4) for (ζ, z) look as follows: for all i, Borel subsets $A \subseteq \mathbb{R}^2_+$ and $t \geq 0$,

$$\zeta_i(A) = \zeta_i(A + (\lambda_i(z)t, t)) + \eta_i \int_0^t \theta_i(A + (\lambda_i(z)s, s)) ds, \tag{6}$$

$$z_i = \zeta_i([\lambda_i(z)t, \infty) \times [t, \infty)) + \eta_i \int_0^t \mathbb{P}\{B_i \geq \lambda_i(z)s, D_i \geq s\} ds. \tag{7}$$

Letting $t \to \infty$ in (6) and (7), we obtain the equations

$$\zeta_i(A) = \eta_i \int_0^\infty \theta_i(A + (\lambda_i(z)s, s))ds, \tag{8}$$

$$z_i = \eta_i \mathbb{E}(B_i/\lambda_i(z) \wedge D_i), \tag{9}$$

which are actually equivalent to (6) and (7).

Thus, we have the closed-form equation (9) for the numeric components of invariant FMS's, and the corresponding measure-valued components are defined by (8). In particular, uniqueness of an invariant FMS is equivalent to uniqueness of a solution to (9).

Definition 2. For an $\mathbb{R}$-valued r.v. X, denote by $\inf X$ the left most point of its support. Recall that the *support* of X is the minimal (in the sense of inclusion) closed interval S such that $\mathbb{P}\{X \in S\} = 1$.

Theorem 2. *Let $1/m_i \geq \inf(D_i/B_i)$ for all i. Then there exists a unique invariant FMS (ζ^*, z^*), and the bandwidth allocation vector $\Lambda(z^*)$ is the unique solution to the optimization problem*

$$\textit{maximize} \quad \sum\nolimits_{i=1}^{I} z_i\, G_i(\Lambda_i) \quad \textit{subject to} \quad A\Lambda \leq C, \quad \Lambda_i \leq \eta_i \mathbb{E}(B_i \wedge m_i D_i) \quad \textit{for all } i, \tag{10}$$

with strictly concave functions $G_i(\cdot)$.

To prove the theorem, we use the approach of Borst *et al.* [4, Lemma 5.2], which is as follows. For now, let z_* be any invariant NFMS As we plug the fixed point equation (9) into the optimization problem (1) for the rate vector $\lambda(z_*)$, the problem (10) follows, which is strictly concave and does not depend on z_*. Hence, $\Lambda(z_*)$ is the same for all invariant points z_*. The assumptions of the theorem guarantee invertibility of the function $\Lambda(\cdot)$, and then uniqueness of $\Lambda(z_*)$ implies uniqueness of z_*.

Note that this method not only proves uniqueness of an invariant FMS, but also suggests a way to compute it (a strictly concave optimization problem can be solved with any desired accuracy in a polynomial time).

Asymptotic bounds for FMS's Here we derive asymptotic bounds for NFMS's that, for a wide class of bandwidth-sharing networks, imply convergence to the invariant NFMS provided it is unique.

Theorem 3. *There exist constants $l, u \in (0, \infty)^I$ such that, for any NFMS $z(\cdot)$,*

$$0 < l_i \leq \underline{\lim}_{t\to\infty} z_i(t) \leq \overline{\lim}_{t\to\infty} z_i(t) \leq u_i \quad \textit{for all } i.$$

These constants satisfy the relations

$$l_i = \eta_i \mathbb{E}(B_i/R_i(l,u) \wedge D_i), \quad u_i = \eta_i \mathbb{E}(B_i/r_i(l,u) \wedge D_i) \quad \textit{for all } i, \tag{11}$$

where the functions $r(\cdot,\cdot)$ and $R(\cdot,\cdot)$ are defined by

$$r_i(x,x') := \inf\nolimits_{x\leq z\leq x'} \lambda_i(z), \quad R_i(x,x') := \sup\nolimits_{x\leq z\leq x'} \lambda_i(z) \quad \textit{for all } i \textit{ and } x \leq x'.$$

Remark 4. There could be more than one pair (l, u) solving (11). The asymptotic bounds l and u for NFMS's given by Theorem 3 form one of such pairs.

We now proceed with the proof of Theorem 3.

Proof. Note that if

$$0 < \tilde{l}_i \leq \underline{\lim}_{t\to\infty} z_i(t) \leq \overline{\lim}_{t\to\infty} z_i(t) \leq \tilde{u}_i \quad \text{for all } i, \tag{12}$$

then also

$$\eta_i \mathbb{E}(B_i/R_i(\tilde{l},\tilde{u}) \wedge D_i) \le \underline{\lim}_{t\to\infty} z_i(t) \le \overline{\lim}_{t\to\infty} z_i(t) \le \eta_i \mathbb{E}(B_i/r_i(\tilde{l},\tilde{u}) \wedge D_i) \quad \text{for all } i. \tag{13}$$

Indeed, by (12), for any $\varepsilon \in (0, \min_{1\le i\le I} \tilde{l}_i)$, there exists a t_ε such that

$$\tilde{l}_i - \varepsilon \le z_i(t) \le \tilde{u}_i + \varepsilon \quad \text{for all } i \text{ and } t \ge t_\varepsilon.$$

Introduce the vectors

$$\tilde{l} - \varepsilon := (\tilde{l}_1 - \varepsilon, \dots, \tilde{l}_I - \varepsilon), \quad \tilde{u} + \varepsilon := (\tilde{u}_1 + \varepsilon, \dots, \tilde{u}_I + \varepsilon).$$

Then

$$r_i(\tilde{l} - \varepsilon, \tilde{u} + \varepsilon)(t-s) \le S_i(z,s,t) \le R_i(\tilde{l} - \varepsilon, \tilde{u} + \varepsilon)(t-s) \quad \text{for } t \ge s \ge t_\varepsilon,$$

which, when plugged into the shifted fluid model equation (5b), implies that

$$\begin{aligned} z_i(t) &\ge \eta_i \int_{t_\varepsilon}^{t} \mathbb{P}\{B_i \ge R_i(\tilde{l} - \varepsilon, \tilde{u} + \varepsilon)(t-s), D_i \ge (t-s)\} ds, \\ z_i(t) &\le \zeta_i(t_\varepsilon)\big([S_i(z,t_\varepsilon,t), \infty) \times [t - t_\varepsilon, \infty)\big) \\ &\quad + \eta_i \int_{t_\varepsilon}^{t} \mathbb{P}\{B_i \ge r_i(\tilde{l} - \varepsilon, \tilde{u} + \varepsilon)(t-s), D_i \ge (t-s)\} ds \quad \text{for } t \ge t_\varepsilon, \end{aligned}$$

where $\zeta(\cdot)$ is the corresponding MVFMS. Taking $t \to \infty$ in the last two inequalities, we obtain

$$\eta_i \mathbb{E}(B_i/R_i(\tilde{l} - \varepsilon, \tilde{u} + \varepsilon) \wedge D_i) \le \underline{\lim}_{t\to\infty} z_i(t) \le \overline{\lim}_{t\to\infty} z_i(t) \le \eta_i \mathbb{E}(B_i/r_i(\tilde{l} - \varepsilon, \tilde{u} + \varepsilon) \wedge D_i),$$

and then (13) follows as $\varepsilon \to 0$.

Now we will iterate (12)–(13). The rate constraints plugged into (4) imply the initial bounds

$$0 < l_i^0 := \eta_i \mathbb{E}(B_i/m_i \wedge D_i) \le \underline{\lim}_{t\to\infty} z_i(t) \le \overline{\lim}_{t\to\infty} z_i(t) \le \eta_i \mathbb{E} D_i =: u_i^0 \quad \text{for all } i,$$

and then (12)–(13) yield the recursive bounds

$$\begin{aligned} l_i^k &:= \eta_i \mathbb{E}(B_i/R_i(l^{k-1}, u^{k-1}) \wedge D_i) \le \underline{\lim}_{t\to\infty} z_i(t), \\ u_i^k &:= \eta_i \mathbb{E}(B_i/r_i(l^{k-1}, u^{k-1}) \wedge D_i) \ \ge \overline{\lim}_{t\to\infty} z_i(t) \quad \text{for all } k \in \mathbb{N} \text{ and } i. \end{aligned} \tag{14}$$

The sequence $\{l^k\}_{k\in\mathbb{N}}$ is non-decreasing and bounded from above by u^0. The sequence $\{u^k\}_{k\in\mathbb{N}}$ is non-increasing and bounded from below by l^0. Hence, there exist the limits $\lim l^k =: l$ and $\lim u^k =: u$. In (14), let $k \to \infty$, then (11) follows.

Note finally that the recursive bounds $\{l^k\}_{k\in\mathbb{N}}$ and $\{u^k\}_{k\in\mathbb{N}}$ as well as their limits l and u do not depend on a particular NFMS. □

Asymptotic stability of an invariant fluid model solution It is realistic to assume that transfer rates in a bandwidth-sharing network decrease as its population grows.

Definition 3. If $z' \ge z \in (0,\infty)^I$ implies $\lambda(z') \le \lambda(z)$, the network is called *monotone*.

For monotone networks, the system of equations (11) decomposes into the fixed point equation (9) for the lower bound l and for the upper bound u, implying the following result.

Corollary 1. *Suppose that the network is monotone and has a unique invariant FMS* (ζ^*, z^*). *Then any FMS* $(\zeta, z)(t) \to (\zeta^*, z^*)$ *as* $t \to \infty$.

(It is easy to see how Theorem 3 implies $z(t) \to z^*$ in Corollary 1. In the Appendix we also show that $z(t) \to z^*$ implies $\zeta(t) \to \zeta^*$.)

Example: single link & multiple classes The sufficient condition for uniqueness of an invariant FMS given by Theorem 2 is sometimes also necessary. Consider, for example, processor

sharing in critical load, that is $J = I = 1$ and (omitting the link and class indices) $\rho = C$. In this case, the fixed point equation (9) looks like

$$z = \eta \mathbb{E}\left(\frac{B}{C/z \wedge m} \wedge D\right),$$

which, for z such that $C/z \leq m$ and $Bz/C \leq D$ a.s., reduces to

$$z = \eta \mathbb{E}(Bz/C \wedge D) = \eta \mathbb{E} Bz/C = \rho z/C = z.$$

I.e. any $z \in [C/m, C \inf D/B]$ is an invariant NFMS. In particular, if $\inf D/B > 1/m$, which violates the assumption of Theorem 2, then there is a continuum of invariant FMS's.

For a single link critically loaded by multiple classes of flows, we have an analogous result, which is more complicated to derive and therefore the proof is postponed to Section 6.

Theorem 4. *Assume that $J = 1$ (in what follows we omit the link index), and that the utility functions are $\mathcal{U}_i(x) = \kappa_i \log x$. If $\sum_{i=1}^{I} \eta_i \mathbb{E}(B_i/m_i \wedge D_i) \neq C$, then there is a unique invariant FMS. Otherwise there might be a continuum if invariant FMS's.*

4 Sequence of stochastic models and fluid limit theorem

In this section we study the asymptotic behavior of the stochastic network described in Section 2 as its global parameters — capacities and arrival rates — grow large, while the characteristics of an individual flow remain of a fixed order. We refer to this scaling as the large capacity regime.

Large capacity scaling To a sequence $\mathcal{R}$ of positive numbers increasing to ∞, we associate a sequence of stochastic models as defined in Section 2. We mark all parameters associated with the r-th model with a superscript r and assume the following:

(A.1) network structure, rate constraints and utility function are the same in all models: $A^r = A$, $m^r = m$ and $\mathcal{U}_i^r(\cdot) = \mathcal{U}_i(\cdot)$ for all i;

(A.2) link capacities grow linearly in r: $C^r = rC$;

(A.3) arrival rates grow linearly in r: $\overline{E}^r(\cdot) := E^r(\cdot)/r \Rightarrow \eta(\cdot)$ as $r \to \infty$, where $\eta(t) := t\eta$ and $\eta \in (0, \infty)^I$;

(A.4) flow sizes and patience times remain of a fixed order: for all i, $(B_i^r, D_i^r) \Rightarrow (B_i, D_i)$, where (B_i, D_i) are $(0, \infty)^I$-valued r.v.'s with distributions θ_i and finite mean values $(1/\mu_i, 1/\nu_i)$, and also $(1/\mu_i^r, 1/\nu_i^r) \to (1/\mu_i, 1/\nu_i)$;

(A.5) the scaled initial configuration converges in distribution to a random vector of finite measures: $\overline{\mathcal{Z}}^r(0) := \mathcal{Z}^r(0)/r \Rightarrow \zeta^0$;

(A.6) the projections $\zeta_i^0(\cdot \times \mathbb{R}_+)$ and $\zeta_i^0(\mathbb{R}_+ \times \cdot)$ are a.s. free of atoms for all i.

Fluid limit theorem In the large capacity regime, the stochastic model defined is Section 2 converges to the fluid model defined in Section 3. More precisely, introduce the scaled state descriptors and population processes

$$\overline{\mathcal{Z}}^r(\cdot) := \mathcal{Z}^r(\cdot)/r, \quad \overline{Z}^r(\cdot) := \langle 1, \overline{\mathcal{Z}}^r(\cdot) \rangle = Z^r(\cdot)/r.$$

Also introduce the scaled versions of the two components of the sate descriptor:

$$\overline{\mathcal{Z}}^{r,\,\mathrm{init}}(\cdot) := \mathcal{Z}^{r,\,\mathrm{init}}(\cdot)/r, \qquad \overline{Z}^{r,\,\mathrm{init}}(\cdot) := \langle 1, \overline{\mathcal{Z}}^{r,\,\mathrm{init}}(\cdot) \rangle = Z^{r,\,\mathrm{init}}(\cdot)/r,$$
$$\overline{\mathcal{Z}}^{r,\,\mathrm{new}}(\cdot) := \mathcal{Z}^{r,\,\mathrm{new}}(\cdot)/r, \qquad \overline{Z}^{r,\,\mathrm{new}}(\cdot) := \langle 1, \overline{\mathcal{Z}}^{r,\,\mathrm{new}}(\cdot) \rangle = Z^{r,\,\mathrm{new}}(\cdot)/r.$$

Remark 5. Let $\lambda(\cdot)$ be the rate allocation function in the unscaled network, then

$$\lambda^r(z) = \lambda(z/r),$$
$$S_i^r(Z^r, s, t) := \int_s^t \lambda_i^r(Z^r(u))du = \int_s^t \lambda_i(\overline{Z}^r(u))du =: S_i(\overline{Z}^r, s, t).$$

We now provide the definition of a fluid limit followed by the main result of this section.

Definition 4. We refer to weak limits along convergent subsequences $\{(\overline{\mathcal{Z}}^q, \overline{Z}^q)(\cdot)\}_{q\in\mathcal{Q}}$, $\mathcal{Q} \subseteq \mathcal{R}$, as *fluid limits*.

Theorem 5. *Under the assumptions* (A.1)–(A.6)*, the sequence* $\{(\overline{\mathcal{Z}}^r, \overline{Z}^r)(\cdot)\}_{r\in\mathcal{R}}$ *is tight in* $\mathbf{D}_{\mathbb{R}_+\to\mathbf{M}^I} \times \mathbf{D}_{\mathbb{R}_+\to\mathbb{R}_+^I}$*, and all fluid limits are a.s. FMS's for the data* (η, θ, ζ^0)*. In particular, if there is a unique FMS* $(\mathcal{Z}, Z)(\cdot)$ *for the data* (η, θ, ζ^0)*, then* $(\overline{\mathcal{Z}}^r, \overline{Z}^r)(\cdot) \Rightarrow (\mathcal{Z}, Z)(\cdot)$ *as* $r \to \infty$*.*

The proof follows in Section 7. To show tightness we adjust the techniques of [13] to the two-dimensional case, since in [13] flows are patient and state descriptors are vectors of measures on $\mathbb{R}_+$. The proof of convergence to FMS's follows the lines of that in [11]. It uses the boundedness of fluid limits away from zero, and the key difference is that in [11] this property is guaranteed by the overload regime, while in our model it holds in any load regime due to the rate constraints.

5 Convergence of stationary distributions

Assume that, in the stochastic model defined in Section 2, the arrival processes are Poisson of rates $\eta_1, \ldots, \eta_I$. Then there exists a unique stationary (and also limiting as $t \to \infty$) distribution of the state descriptor $\mathcal{Z}(\cdot)$. Indeed, without loss of generality, there are i.i.d. r.v.'s $\{\widetilde{D}_{ik}\}_{k\in\mathbb{N},1\le i\le I}$ distributed as $\max_{1\le i\le I} D_i$ and such that a.s. $D_{ik} \le \widetilde{D}_{ik}$ for all k and i. Then the total population $\sum_{i=1}^I Z_i(\cdot)$ of the network is a.s. and within the whole time horizon $\mathbb{R}_+$ bounded from above by the length of the $M/G/\infty$ queue with the following parameters. At time $t = 0$, there are $\sum_{i=1}^I Z_i(0)$ customers in the queue whose service times are patience times of initial flows in the network. The input process for the queue is the composition of those for the network, and hence is Poisson of rate $\sum_{i=1}^I \eta_i$. Service times of new customers in the queue are drawn from the sequence $\{\widetilde{D}_{ik}\}_{k\in\mathbb{N},1\le i\le I}$ of upper bounds for patience times of new flows in the network. As any other $M/G/\infty$ queue, the defined queue is regenerative. The instants when a customer enters the empty queue form an embedded renewal process whose cycle length is non-lattice and has a finite mean value $\exp(\sum_{i=1}^I \eta_i\, \mathbb{E}\widetilde{D}_{11})/\sum_{i=1}^I \eta_i$. With respect to this renewal process , the state descriptor $\mathcal{Z}(\cdot)$ is also regenerative. Then, by [1, Chapter V.I, Theorem 1.2], there exists a limiting distribution for $\mathcal{Z}(\cdot)$.

Now consider a sequence of stochastic models as defined in Section 2 that satisfies the assumptions (A.1), (A.2), (A.4) (see Section 4) and

(A′.3) the input processes $E_1^r(\cdot), \ldots, E_I^r(\cdot)$ are independent Poisson processes of rates $\eta_1^r, \ldots, \eta_I^r$, and $\eta^r/r \to \eta \in (0,\infty)^I$ as $r \to \infty$.

Let $\mathcal{Y}^r$ have the stationary distribution of $\mathcal{Z}^r(\cdot)$ and put $Y^r := \langle 1, \mathcal{Y}^r\rangle$. Introduce also the scaled versions

$$\overline{\mathcal{Y}}^r := \mathcal{Y}^r/r, \quad \overline{Y}^r := \langle 1, \overline{\mathcal{Y}}^r\rangle = Y^r/r.$$

We now have the following result.

Theorem 6. *Under the assumptions* (A.1), (A.2), (A′.3) *and* (A.4)*, the sequence* $\{(\overline{\mathcal{Y}}^r, \overline{Y}^r)\}_{r\in\mathcal{R}}$ *is tight, and any weak limit point* $(\mathcal{Y}, Y)$ *is a weak invariant FMS, i.e. there exists a stationary FMS* $(\mathcal{Z}, Z)(t) \stackrel{d}{=} (\mathcal{Y}, Y)$*,* $t \ge 0$*. In particular, by Corollary 1, if the network is monotone and has a unique invariant FMS* (ζ_*, z_*)*, then* $(\overline{\mathcal{Y}}^r, \overline{Y}^r) \Rightarrow (\zeta_*, z_*)$ *as* $r \to \infty$*.*

The general strategy of the proof is adopted from [17, Theorem 3.3]: we check that any convergent subsequence of initial conditions $\{\overline{\mathcal{Z}}^q(0) \stackrel{\mathrm{d}}{=} \overline{\mathcal{Y}}^q\}_{q\in\mathcal{Q}}$, $\mathcal{Q}\subseteq\mathcal{R}$, $\overline{\mathcal{Y}}^q \Rightarrow \mathcal{Y}$, satisfies the assumptions of the fluid limit theorem (we only need to check (A.6)). Then the corresponding subsequence $\{\overline{\mathcal{Z}}^q(\cdot)\}_{q\in\mathcal{Q}}$ of the scaled state descriptors converges to an MVFMS that is stationary (i.e. $\mathcal{Y}$ is a weak invariant MVFMS) since all $\overline{\mathcal{Z}}^q(\cdot)$ are stationary.

The techniques we use to implement this strategy are different from the techniques of [17], though. Our key instruments for establishing tightness are $M/G/\infty$ bounds, see Section 8. Below we present an elegant proof of (A.6) for weak limit points of $\{\overline{\mathcal{Y}}^r\}_{r\in\mathcal{R}}$.

Lemma 2. *Any weak limit point $\mathcal{Y}$ of $\{\overline{\mathcal{Y}}^r\}_{r\in\mathcal{R}}$ has both projections $\mathcal{Y}(\cdot\times\mathbb{R}_+)$ and $\mathcal{Y}(\mathbb{R}_+\times\cdot)$ a.s. free of atoms.*

Proof. The key idea is the following. Consider the network in its stationary regime. Then, on one hand, it always has the same distribution, and on the other hand, all initial flows are gone at some point, and newly arriving flows do not accumulate along horizontal and vertical lines.

Let $\mathcal{Y}$ be the weak limit along a subsequence $\{\overline{\mathcal{Y}}^q\}_{q\in\mathcal{Q}}$, and run the q-th network starting from $\overline{\mathcal{Z}}^q(0) \stackrel{\mathrm{d}}{=} \overline{\mathcal{Y}}^q$. By [11, Lemma 6.2], it suffices to show that, for any $\delta>0$ and $\varepsilon>0$, there exists an $a>0$ such that

$$\underline{\lim}_{q\to\infty}\mathbb{P}^q\{\sup_{x\in\mathbb{R}_+}\|\overline{\mathcal{Y}}^q(H_x^{x+a})\|\vee\|\overline{\mathcal{Y}}^q(V_x^{x+a})\|\le\delta\}\ge 1-\varepsilon, \tag{15}$$

where $H_a^b := \mathbb{R}_+\times[a,b]$ and $V_a^b := [a,b]\times\mathbb{R}_+$ for all $b\ge a\ge 0$.

First we estimate the time when there is only a few (when scaled) initial flows left. The initial flows whose initial patience times are less than t are already gone at time t. Then Lemma 16 (see Section 8) implies that (recall that $\Pi(\lambda)$ stands for the Poisson distribution with parameter λ)

$$\begin{aligned}\overline{Z}_i^{q,\,\mathrm{init}}(t) \le \overline{\mathcal{Z}}_i^q(0)(\mathbb{R}_+\times[t,\infty)) &\stackrel{\mathrm{d}}{=} \overline{\mathcal{Y}}_i^q(\mathbb{R}_+\times[t,\infty))\\ &\le_{\mathrm{st}} \frac{1}{q}\Pi(\eta_i^q\int_t^\infty \mathbb{P}^q\{D_i^q>y\}dy) \Rightarrow \eta_i\int_t^\infty\mathbb{P}\{D_i>y\}dy.\end{aligned}$$

Take T such that $\max_{1\le i\le I}\eta_i\int_t^\infty\mathbb{P}\{D_i>y\}dy<\delta/2$, then

$$\lim_{q\to\infty}\mathbb{P}^r\{\|\overline{Z}^{q,\,\mathrm{init}}(T)\|\le\delta/2\}=1.$$

Now, in Lemma 11 (see Section 7), we prove that newly arriving customers do not accumulate in thin horizontal and vertical strips, i.e. there exists an $a>0$ such that

$$\underline{\lim}_{q\to\infty}\mathbb{P}^q\{\sup_{t\in[0,T]}\sup_{x\in\mathbb{R}_+}\|\overline{\mathcal{Z}}^{q,\,\mathrm{new}}(t)(H_x^{x+a})\|\vee\|\overline{\mathcal{Z}}^{q,\,\mathrm{new}}(t)(V_x^{x+a})\|\le\delta/2\}\ge 1-\varepsilon.$$

Finally, because of stationarity of $\mathcal{Y}^q$,

$$\begin{aligned}&\mathbb{P}^q\{\sup_{x\in\mathbb{R}_+}\|\overline{\mathcal{Y}}^q(H_x^{x+a})\|\vee\|\overline{\mathcal{Y}}^q(V_x^{x+a})\|\le\delta\}\\ =&\mathbb{P}^q\{\sup_{x\in\mathbb{R}_+}\|\overline{\mathcal{Z}}^q(T)(H_x^{x+a})\|\vee\|\overline{\mathcal{Z}}^q(T)(V_x^{x+a})\|\le\delta\}\\ \ge&\mathbb{P}^q\{\|\overline{Z}^{q,\,\mathrm{init}}(T)\|\le\delta/2, \sup_{x\in\mathbb{R}_+}\|\overline{\mathcal{Z}}^{q,\,\mathrm{init}}(T)(H_x^{x+a})\|\vee\|\overline{\mathcal{Z}}^{q,\,\mathrm{init}}(T)(V_x^{x+a})\|\le\delta/2\},\end{aligned}$$

which implies (15) by the choice of T and a. □

6 Proof of fluid model properties

Here we prove the results of Section 3.

6.1 Proof of Theorem 1

In the proof of Theorem 1, we exploit boundedness of NFMS's away from zero and in the norm (see Lemma 3), and Lipschitz continuity of MVFMS's in the first coordinate (see Lemma 5). We also use the auxiliary Lemma 4.

Lemma 3. *Let $z(\cdot)$ be an NFMS Then $\sup_{t\geq 0}\|z(t)\| \leq \|z(0)\| + \|\sigma\| < \infty$ and, for any $\delta > 0$, $\inf_{t\geq\delta}\min_{1\leq i\leq I} z_i(t) > 0$. In particular, if $z_i(0) > 0$, then $\inf_{t\geq 0} z_i(t) > 0$.*

Proof. By the rate constraints, $S_i(z,s,t) \leq m_i(t-s)$ for all $s \leq t$, which, when plugged into the fluid model equation (4), implies the following lower bound: $z_i(t) \geq \eta_i \int_0^t \mathbb{P}\{B_i/m_i \wedge D_i \geq s\}ds$. Since $f_i(s) := \mathbb{P}\{B_i/m_i \wedge D_i \geq s\} \uparrow \mathbb{P}\{B_i/m_i \wedge D_i > 0\} = 1$ as $s \downarrow 0$, in a small enough interval $[0,\varepsilon]$, $f_i(\cdot) \geq 1/2$. Then, for $t \geq \delta$, $z_i(t) \geq \eta_i \int_0^{\delta\wedge\varepsilon} f_i(s)ds \geq (\delta\wedge\varepsilon)/2$. The upper bound follows from (4) directly: $z_i(t) \leq z_i(0) + \eta_i \int_0^t \mathbb{P}\{D_i \geq s\}ds \uparrow z_i(0) + \sigma_i$ as $t \uparrow \infty$. □

Lemma 4. *For an $\mathbb{R}$-valued r.v. ξ and $x \leq x'$, $\int_{\mathbb{R}} \mathbb{P}\{u + x \leq \xi \leq u + x'\}du \leq x' - x$.*

See the Appendix for the proof.

Lemma 5. *Under assumption* (ii) *of Theorem 1, any MVFMS $\zeta(\cdot)$ at any time $t \geq 0$ has a Lipschitz continuous first projection, i.e. there exists a constant $L(\zeta,t) \in (0,\infty)$ such that for all i, $x < x'$ and y,*

$$\zeta_i(t)([x,x'] \times [y,\infty)) \leq L(\zeta,t)(x' - x).$$

Proof. For an FMS $(\zeta, z)(\cdot)$, for all i, $t \geq 0$, $x < x'$ and y,

$$\zeta_i(t)([x,x'] \times [y,\infty)) \leq f_i(x,x',y) + \eta_i g_i(x,x',y),$$

where

$$\begin{aligned} f_i(x,x',y) &:= \zeta_i^0([x + S_i(z,0,t), x' + S_i(z,0,t)] \times [y + t, \infty)), \\ g_i(x,x',y) &:= \int_0^t \mathbb{P}\{x + S_i(z,s,t) \leq B_i \leq x' + S_i(z,s,t)\}ds. \end{aligned} \tag{16}$$

By Lipschitz continuity of the initial condition, $f_i(x,x',y) \leq L(x' - x)$. In (16), change the variable of integration for $v = V(s) := S_i(z,s,t)$. Then

$$g_i(x,x',y) = \int_0^{S_i(z,0,t)} \mathbb{P}\{x + v \leq B_i \leq x' + v\}/\lambda_i(z(V^{-1}(v)))dv \leq M(\zeta,t)(x' - x),$$

where $M(\zeta,t) := \sup_{s\in[0,t]}\max_{1\leq i\leq I} 1/\lambda_i(z(s))$. By Lemma 3, the functions $1/\lambda_i(z(\cdot))$ are continuous in $[0,t]$. Hence $M(\zeta,t)$ is finite and the first projection of $\zeta(t)$ is Lipschitz continuous with the constant $L(\zeta,t) := L + \|\eta\|M(\zeta,t)$. □

Now we are in a position to prove Theorem 1.

Proof of Theorem 1. Let $(\zeta^1, z^1)(\cdot)$ and $(\zeta^2, z^2)(\cdot)$ be two FMS's for the data (η, θ, ζ^0).

(i) We show that the two FMS's coincide in an interval $[0,\delta]$. We check that $z^1(\delta) = z^2(\delta) \in (0,\infty)^I$ and that the first projection of $\zeta^1(\delta) = \zeta^2(\delta)$ is Lipschitz continuous. Then, by Remark 2 and the second part of the theorem, the two FMS's coincide everywhere.

Note that, for a vector $z \in (0,\infty)^I$ of a small enough norm, $\lambda_i(z) = m_i$ for all i. Lemma 3 and the fluid model equation (4) imply that $0 < z_i^1(t), z_i^2(t) \leq \eta_i t$ for all i and $t > 0$. Then, for all i and $s,t \in [0,\delta]$, where δ is small enough,

$$S_i(z^1,s,t) = S_i(z^2,s,t) = m_i(t-s). \tag{17}$$

Plugging (17) into (4), we obtain, for $t \in [0, \delta]$ and all i,

$$z_i^1(t) = z_i^2(t) = \eta_i \int_0^t \mathbb{P}\{B_i/m_i \wedge D_i \geq s\} ds.$$

By Remark 3, $\zeta^1(\cdot)$ and $\zeta^2(\cdot)$ coincide in $[0, \delta]$, too. Lipschitz continuity of the first projection of $\zeta^1(\delta) = \zeta^2(\delta)$ follows as we plug (17) into the fluid model equation (3) (recall that it is valid for all Borel sets): for all i, $x < x'$ and y,

$$\begin{aligned}
\zeta_i^j(\delta)([x, x'] \times [y, \infty)) &= \eta_i \int_0^\delta \mathbb{P}\{x + m_i s \leq B_i \leq x' + m_i s, D_i \geq y + s\} ds \\
&\leq \eta_i \int_0^\delta \mathbb{P}\{x/m_i + s \leq B_i/m_i \leq x'/m_i + s\} ds \\
&\leq \eta_i (x' - x)/m_i, \quad j = 1, 2,
\end{aligned}$$

where the last inequality holds by Lemma 4.

(ii) Suppose that the two FMS's are different, that is $t_* := \inf\{t > 0 \colon z^1(t) \neq z^2(t)\} < \infty$.

Without loss of generality we may assume that $t_* = 0$. Indeed, otherwise we can consider the time-shifted FMS's $(\zeta^j, z^j)(t_* + \cdot)$, $j = 1, 2$. By Lemmas 3 and 5, they start from $z^1(t_*) = z^2(t_*) \in (0, \infty)^I$ and $\zeta^1(t_*) = \zeta^2(t_*)$ with a Lipschitz continuous first projection.

By Lemma 3, the two NFMS never leave a compact set $[\delta, \Delta]^I \subset (0, \infty)^I$. Since the rate functions $\lambda_i(z)$ are Lipschitz continuous in such sets, there exists a constant $K \in (0, \infty)$ such that, for all i and $s \leq t$,

$$|S_i(z^1, s, t) - S_i(z^2, s, t)| \leq Kt \sup_{s \in [0,t]} \|z^1(s) - z^2(s)\| =: Kt\varepsilon(t).$$

Then, by Lipschitz continuity of the initial condition, we have, for all i and $t \geq 0$,

$$|z_i^1(t) - z_i^2(t)| \leq LKt\varepsilon(t) + \eta_i \int_0^t \mathbb{P}\{S_i(z^1, s, t) - Kt\varepsilon(t) \leq B_i \leq S_i(z^1, s, t) + Kt\varepsilon(t)\} ds.$$

In the last equation, change the variable of integration for $v = S_i(z^1, s, t)$ (*cf.* the proof of Lemma 5) and put $M = \sup_{z \in [\delta, \Delta]^I} \max_{1 \leq i \leq I} 1/\lambda_i(z)$. Then

$$|z_i^1(t) - z_i^2(t)| \leq LKt\varepsilon(t) + \eta_i M 2Kt\varepsilon(t) \text{ for all } i \quad \text{and} \quad \varepsilon(t) \leq (L + 2\|\eta\| M) Kt\varepsilon(t),$$

which implies that $\varepsilon(t) = 0$ for small enough t, and we arrive at a contradiction with $t_* = 0$. □

6.2 Proof of Theorem 2

Multiplying the coordinates of (9) by the corresponding rates $\lambda_i(z)$, we obtain the equivalent fixed point equation

$$\Lambda_i(z) = g_i(\lambda_i(z)) \quad \text{for all } i, \tag{18}$$

where

$$g_i(x) := \eta_i \mathbb{E}(B_i \wedge x D_i), \quad x \geq 0.$$

We first discuss some properties of the functions $g_i(\cdot)$ in the auxiliary Lemmas 6 and 7, and then proceed with the proof of the theorem.

Lemma 6. *The function $g_i(\cdot)$ is continuous. Also $g_i(\cdot)$ is strictly increasing in $[0, \alpha_i)$ and constant in $\mathbb{R}_+ \setminus [0, \alpha_i)$, where*

$$\alpha_i := \inf\{x \colon g_i(x) = \rho_i\} > 0,$$

and infimum over the empty set is defined to be ∞.

Proof. Continuity of $g_i(\cdot)$ follows by the dominated convergence theorem.

The situation $\alpha_i = 0$ is not possible since in that case $g_i(x) = \rho_i$ for all $x > 0$ by the definition of α_i. But $g_i(\cdot)$ is continuous and $g_i(x) \to g_i(0) = 0$ as $x \to 0$.

If $\alpha_i < \infty$, then, again by the definition of α_i and continuity of $g_i(\cdot)$, we have $g_i(x) = \rho_i$ for all $x \geq \alpha_i$.

It is left to check that $g_i(\cdot)$ is strictly increasing in $[0, \alpha_i)$. Assume that $0 \leq x < y < \alpha_i$, but $g_i(x) = g_i(y)$. Then

$$\begin{aligned} 0 = g_i(y)/\eta_i - g_i(x)/\eta_i =& \mathbb{E}B_i\mathbb{I}_{\{B_i \leq xD_i\}} + \mathbb{E}B_i\mathbb{I}_{\{xD_i < B_i \leq yD_i\}} + \mathbb{E}yD_i\mathbb{I}_{\{B_i > yD_i\}} \\ & - \mathbb{E}B_i\mathbb{I}_{\{B_i \leq xD_i\}} - \mathbb{E}xD_i\mathbb{I}_{\{xD_i < B_i \leq yD_i\}} - \mathbb{E}xD_i\mathbb{I}_{\{B_i > yD_i\}}) \\ =& \mathbb{E}\underbrace{(B_i - xD_i)\mathbb{I}_{\{xD_i < B_i \leq yD_i\}}}_{=:\,X} + (y-x)\mathbb{E}\underbrace{D_i\mathbb{I}_{\{B_i > yD_i\}}}_{=:\,Y}, \end{aligned}$$

where the r.v.'s X and Y are non-negative, so they must a.s. equal zero. In particular, we have $B_i \leq yD_i$ and $g_i(y) = \rho_i$, which contradicts the definition of α_i since $y < \alpha_i$. □

The stabilization points α_i of the functions $g_i(\cdot)$ are related with the r.v.'s (B_i, D_i) in the following way.

Lemma 7. *If $\alpha_i < \infty$, then $\inf D_i/B_i = 1/\alpha_i$. If $\alpha_i = \infty$, then $\inf D_i/B_i = 0$.*

Proof. First assume $\alpha_i < \infty$. Rewrite the relation $g_i(x) = \rho_i$ as $\mathbb{E}B_i(1 - (1 \wedge xD_i/B_i)) = 0$, which, for $x > 0$, is equivalent to $D_i/B_i \geq 1/x$ a.s. Hence $\alpha_i = \inf\{x > 0\colon D_i/B_i \geq 1/x \text{ a.s.}\}$ and $1/\alpha_i = \sup\{y > 0\colon D_i/B_i \geq y \text{ a.s.}\}$. In the right-hand side of the latter equation we see the definition of $\inf D_i/B_i$.

Now consider the case $\alpha_i = \infty$. Assume that $\inf D_i/B_i = y > 0$, then $D_i/y \geq B_i$ a.s. and $g_i(1/y) = \rho_i$. On the other hand, since $\alpha_i = \infty$, there is no $x > 0$ such that $g_i(x) = \rho_i$. Hence $y = 0$. □

Having established the above properties of $g_i(\cdot)$'s, we now can prove Theorem 2 by means of a technique developed by Borst *et al.* [4, Lemma 5.2].

Proof of Theorem 2. First show uniqueness. Let $z^* \in (0, \infty)^I$ be an invariant NFMS. Recall that $\lambda(z^*)$ is the unique optimal solution for the concave optimization problem (1). The necessary and sufficient conditions for that are given by the Karush-Kuhn-Tucker (KKT) theorem (see e.g. [3, Theorem 3.1]): there exist $p \in \mathbb{R}_+^J$ and $q \in \mathbb{R}_+^I$ such that

$$\mathcal{U}_i'(\lambda_i(z^*)) = \sum\nolimits_{j=1}^{J} A_{ji}p_j + q_i \quad \text{for all } i, \tag{19a}$$

$$p_j(\sum\nolimits_{i=1}^{I} A_{ji}\Lambda_i(z^*) - C_j) = 0 \quad \text{for all } j, \tag{19b}$$

$$q_i(\lambda_i(z^*) - m_i) = 0 \quad \text{for all } i. \tag{19c}$$

By the assumptions of the theorem and Lemmas 6 and 7, the functions $g_i(\cdot)$ are strictly increasing in the intervals $[0, m_i]$, which implies two things (see also Fig. 2). First, the fixed point equation (18) can be rewritten as $\lambda_i(z^*) = g_i^{-1}(\Lambda_i(z^*))$ for all i, and we plug that into (19a). Second, the second multiplier in (19c) is zero if and only if $g_i(\lambda_i(z^*)) = g_i(m_i)$, and that, by (18), is equivalent to $\Lambda_i(z^*) = g_i(m_i)$. Hence, $\Lambda(z^*)$ satisfies

$$\mathcal{U}_i'(g_i^{-1}(\Lambda_i(z^*)) = \sum\nolimits_{j=1}^{J} A_{ji}p_j + q_i \quad \text{for all } i, \tag{20a}$$

$$p_j(\sum\nolimits_{i=1}^{I} A_{ji}\Lambda_i(z^*) - C_j) = 0 \quad \text{for all } j, \tag{20b}$$

$$q_i(\Lambda_i(z^*) - g_i(m_i)) = 0 \quad \text{for all } i. \tag{20c}$$

Now note that the last three equations form the KKT conditions for another optimization problem. Indeed, take functions $\tilde{g}_i(\cdot)$ that are continuous and strictly increasing in $\mathbb{R}_+$ and coincide with $g_i(\cdot)$ in $[0, m_i]$. Also take functions $G_i(\cdot)$ such that $G_i'(\cdot) = \mathcal{U}_i'(\tilde{g}_i^{-1}(\cdot))$ in $(0, \infty)$. Then (20) gives necessary and sufficient conditions for $\Lambda(z^*)$ to

$$\text{maximize} \quad \sum\nolimits_{i=1}^{I} z_i \, G_i(\Lambda_i) \quad \text{subject to} \quad A\Lambda \le C, \quad \Lambda_i \le g_i(m_i) \quad \text{for all } i. \tag{21}$$

Since the functions $\mathcal{U}_i(\cdot)$ are strictly concave, their derivatives $\mathcal{U}_i'(\cdot)$ are strictly decreasing. Then $G_i'(\cdot)$ are strictly decreasing and, equivalently, $G_i(\cdot)$ are strictly concave, which implies that $\Lambda(z^*)$ is actually the unique solution to (21). Then the invariant NFMS z^* must be unique, too, $z_i^* = \Lambda_i(z^*)/g_i^{-1}(\Lambda_i(z^*))$ for all i.

The existence result follows similarly. There exists a unique optimal solution Λ^* to (21) and it satisfies the KKT conditions (20). Put $\lambda_i^* = g_i^{-1}(\Lambda_i^*)$ for all i. Then λ^* and Λ^* satisfy the KKT conditions (19), i.e., for the vector z^* with $z_i^* := \Lambda_i^*/\lambda_i^*$, we have $\lambda^* = \lambda(z^*)$ and $\Lambda^* = \Lambda(z^*)$. Plugging the last two relations into the definition of λ^*, we get the fixed point equation. □

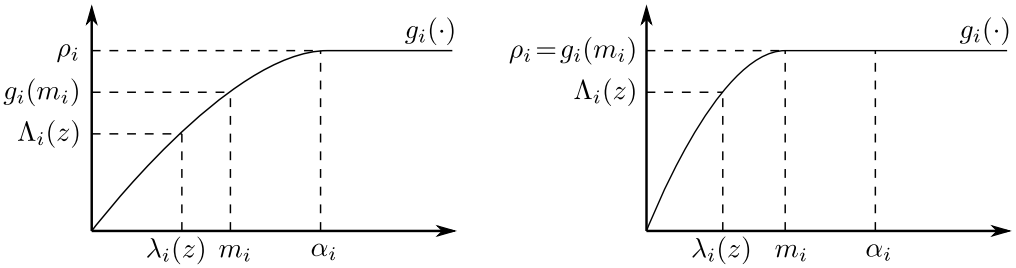

Figure 2: Graph of the function $g_i(\cdot)$ in the two possible cases: when $m_i \le \alpha_i$ (left) and when $m_i > \alpha_i$ (right); z is a invariant NFMS.

6.3 Proof of Theorem 4

The fixed point equation (18) and the monotonicity of the functions $g_i(\cdot)$ imply that the bandwidth class i gets in an equilibrium is at most $g_i(m_i)$. Therefore, we refer to the scenarios $\sum_{i=1}^{I} g_i(m_i) < C$, $\sum_{i=1}^{I} g_i(m_i) = C$ and $\sum_{i=1}^{I} g_i(m_i) > C$ as underloaded, critically loaded and overloaded, respectively. Below we calculate the invariant NFMS's in the three cases.

Summing up (9), the KKT conditions for (1) and the capacity and rate constraints, a $z \in (0, \infty)^I$ is an invariant NFMS if and only if there exist $p \in \mathbb{R}_+$ and $q \in \mathbb{R}_+^I$ such that (we omit the argument of the rates $\lambda_i(z)$ and bandwidth allocations $\Lambda_i(z)$)

$$\Lambda_i = g_i(\lambda_i) \quad \text{for all } i, \tag{22a}$$

$$\kappa_i/\lambda_i = p + q_i \quad \text{for all } i, \tag{22b}$$

$$p(\sum\nolimits_{i=1}^{I} \Lambda_i - C) = 0 \tag{22c}$$

$$q_i(\lambda_i - m_i) = 0 \quad \text{for all } i, \tag{22d}$$

$$\sum\nolimits_{i=1}^{I} \Lambda_i \le C, \tag{22e}$$

$$\lambda_i \le m_i \quad \text{for all } i. \tag{22f}$$

Underload In this case, there is no interaction between the classes, they do not compete but all get the maximum rate allowed. Indeed, (22c) and (22b) imply that $p = 0$ and all $q_i > 0$. Then, by (22d) and (22a), all $\lambda_i = m_i$ and all $\Lambda_i = g_i(m_i)$. Hence, there is a unique invariant NFMS given by

$$z_i = g_i(m_i)/m_i \quad \text{for all } i.$$

Critical load First note that

$$\Lambda_i = g_i(m_i) \quad \text{for all } i. \tag{23}$$

Indeed, there are two possibilities: either $p = 0 \overset{(22b)}{\Rightarrow}$ all $q_i > 0 \overset{(22d)}{\Rightarrow}$ all $\lambda_i = m_i \overset{(22a)}{\Rightarrow}$ (23), or $p > 0 \overset{(22c)}{\Rightarrow} \sum_{i=1}^I \Lambda_i = C \Rightarrow$ (23), where the last implication is due to $\Lambda_i \le g_i(m_i)$ and $\sum_{i=1}^I g_i(m_i) = C$. Recall from Lemma 7 that

$$\alpha_i := \inf\{x\colon g_i(x) = \rho_i\} = 1/\inf(D_i/B_i).$$

By (23), the relations (22a) and (22f) are equivalent to $m_i \wedge \alpha_i \le \lambda_i \le m_i$ (see Fig. 2). Hence, (22) reduces to

$$\kappa_i/\lambda_i = p + q_i, \tag{24a}$$

$$q_i(\lambda_i - m_i) = 0, \tag{24b}$$

$$m_i \wedge \alpha_i \le \lambda_i \le m_i. \tag{24c}$$

Let

$$\mathcal{I}_{\text{crit}} := \{i\colon m_i \le \alpha_i\}.$$

For $i \in \mathcal{I}_{\text{crit}}$, by (24c), we have $\lambda_i = m_i$ and $z_i = g_i(m_i)/m_i$. Then (24b) is satisfied, and (24a) implies that $p \le \kappa_i/m_i$.

Now divide $\{1, \dots, I\} \setminus \mathcal{I}_{\text{crit}}$ into two subsets $\mathcal{I}_1 \cap \mathcal{I}_2 = \varnothing$. For $i \in \mathcal{I}_1$, put $\lambda_i = m_i$, then (as for $i \in \mathcal{I}_{\text{crit}}$) $z_i = g_i(m_i)/m_i$, (24b) is satisfied, and (24a) implies that $p \le \kappa_i/m_i$. For $i \in \mathcal{I}_2$, assume $\alpha_i \le \lambda_i < m_i$. Then $q_i = 0$ by (24b), $\kappa_i/\lambda_i = p$ by (24a), and $\kappa_i/m_i < p \le \kappa_i/m_i$. Also $z = g_i(m_i)p/\kappa_i$.

Summing up everything said above, the set of invariant NFMS's is given by

$$S_z := \bigcup\nolimits_{\mathcal{I} \supseteq \mathcal{I}_{\text{crit}}} \{z\colon z_i = g_i(m_i)/m_i \text{ for } i \in \mathcal{I} \text{ and } z_i = g_i(m_i)p/\kappa_i \text{ for } i \notin \mathcal{I}, \\ \text{where } p \in (\max_{i\notin\mathcal{I}} \kappa_i/m_i, \min_{i\in\mathcal{I}} \kappa_i/m_i \wedge \min_{i\notin\mathcal{I}} \kappa_i/\alpha_i]\}.$$

Equivalent descriptions of S_z are

$$S_z = \{z\colon z_i = g_i(m_i)/m_i \text{ if } p \le \kappa_i/m_i \text{ and } z_i = g_i(m_i)p/\kappa_i \text{ if } p > \kappa_i/m_i, \quad p \in S_p\},$$
$$S_p := (0, \min_{i\in\mathcal{I}_{\text{crit}}} \kappa_i/m_i \wedge \min_{i\notin\mathcal{I}_{\text{crit}}} \kappa_i/\alpha_i] = (0, \min_{1\le i\le I} \kappa_i/(m_i \wedge \alpha_i)],$$

and

$$S_z = \{z\colon z_i = g_i(m_i)/(m_i \wedge \kappa_i x), \quad x \in S_x\},$$
$$S_x := [\max_{1\le i\le I}(m_i \wedge \alpha_i)/\kappa_i, \infty).$$

We now apply the last formula in a couple of simple examples.

Example 1. If $m_i \le \alpha_i$ for all i, then $S_x = [\max_{1\le i\le I} m_i/\kappa_i, \infty)$, and $\kappa_i x \ge m_i$ for all $x \in S_x$ and all i. Hence, there is a unique invariant NFMS given by $z_i = g_i(m_i)/m_i$ for all i, which agrees with Theorem (2).

Example 2. If $m_1 > \alpha_1$, $m_i \le \alpha_i$ for $i \ne 1$ and $\alpha_1/\kappa_1 \ge \max_{i\ne 1} m_i/\kappa_i$, then, for any $\lambda_1 \in [\alpha_1, m_1]$, $z = (g_1(m_1)/\lambda_1, g_2(m_2)/m_2, \dots, g_I(m_I)/m_I)$ is an invariant NFMS.

Overload In this situation, by the capacity constraint (22e), at least one class of flows does not receive the maximum service, i.e. at least one $\Lambda_i < g_i(m_i)$. We first find out which classes get the maximum service and which do not, and then calculate the unique invariant NFMS.

Who gets the maximum service. Since at least one $\Lambda_i < g_i(m_i)$, at least one $\lambda_i < m_i \wedge \alpha_i$ (see Fig. 2). Then (22d), (22b) and (22c) imply that at least one $q_i = 0$, $p > 0$ and $\sum_{i=1}^I \Lambda_i = C$. At this point, we can equivalently rewrite (22) as follows: there exist $x > 0$ and $\varepsilon \in \mathbb{R}_+^I$ such

that (the functions $\tilde{g}_i(\cdot)$ are introduced below)

$$\Lambda_i = g_i(\lambda_i) \Leftrightarrow \Lambda_i = \tilde{g}_i(\lambda_i), \tag{25a}$$

$$\sum\nolimits_{i=1}^{I} g_i(\lambda_i) = C \Leftrightarrow \sum\nolimits_{i=1}^{I} \tilde{g}_i(\lambda_i) = C \tag{25b}$$

$$\lambda_i = \kappa_i(x - \varepsilon_i), \tag{25c}$$

$$\varepsilon_i(\lambda_i - m_i) = 0, \tag{25d}$$

$$\lambda_i \le m_i. \tag{25e}$$

For all i and $x \ge 0$, put

$$\tilde{g}_i(x) := g_i(m_i \wedge x).$$

By the rate constraints (25e), in (25), we can equivalently replace $g_i(\cdot)$ by $\tilde{g}_i(\cdot)$.

If $\lambda_i < m_i$, then, by (25d) and (25c), $\varepsilon_i = 0$ and $\lambda_i = \kappa_i x$, and hence

$$\tilde{g}_i(\lambda_i) = \tilde{g}_i(\kappa_i x). \tag{26}$$

If $\lambda_i = m_i$, then, by (25c), $\kappa_i x \ge m_i$ and $\tilde{g}_i(\kappa_i x) = g_i(m_i)$, and, again, (26) holds.

Plugging (26) into (25b), we get

$$\sum\nolimits_{i=1}^{I} \tilde{g}_i(\kappa_i x) = C. \tag{27}$$

The function $\tilde{g}(x) := \sum_{i=1}^{I} \tilde{g}_i(\kappa_i x)$ is continuous everywhere, strictly increasing in the interval

$$0 \le x \le \max\nolimits_{1 \le i \le I}(m_i \wedge \alpha_i)/\kappa_i =: x_0$$

and constant for $x \ge x_0$, and also $\tilde{g}(0) = 0$ and $\tilde{g}(x_0) = \sum_{i=1}^{I} g_i(m_i) > C$, which implies that there exists a unique x solving (27) and $x \in (0, x_0)$.

By (25a) and (26), $\Lambda_i = \tilde{g}_i(\kappa_i x)$. Then (see Fig. 3) $\Lambda_i = g_i(m_i)$ if $(m_i \wedge \alpha_i)/\kappa_i \le x$ and $\Lambda_i < g_i(m_i)$ if $(m_i \wedge \alpha_i)/\kappa_i > x$. Hence, the set of classes that get the maximum service is

$$\mathcal{I}_{\text{over}} := \{i\colon (m_i \wedge \alpha_i)/\kappa_i \le x\}. \tag{28}$$

Invariant NFMS. For $i \notin \mathcal{I}_{\text{over}}$, $\Lambda_i = g_i(\lambda_i) < g_i(m_i)$, which implies that (see Fig. 2) $\lambda_i < m_i \wedge \alpha_i$. Then, by (25d) and (25c), $\varepsilon_i = 0$ and $\lambda_i = \kappa_i x$ (meeting the rate constraint (25e)), and $z_i = \Lambda_i/\lambda_i = g_i(\kappa_i x)/(\kappa_i x)$.

For $i \in \mathcal{I}_{\text{over}}$, consider the two possible cases: $\kappa_i x < m_i$ and $\kappa_i x \ge m_i$. If $\kappa_i x < m_i$, then, by (25c) and (25d), $\lambda_i \le \kappa_i x < m_i$ and $\varepsilon_i = 0$, and, again by (25c), $\lambda_i = \kappa_i x$. If $\kappa_i x \ge m_i$, then $\lambda_i = m_i$ because otherwise we would arrive at a contradiction: $\lambda_i < m_i \overset{(25d)}{\Rightarrow} \varepsilon_i = 0 \overset{(25c)}{\Rightarrow} \lambda_i = \kappa_i x \ge m_i$. Hence, for $i \in \mathcal{I}_{\text{over}}$, $\lambda_i = m_i \wedge \kappa_i x$ and $z_i = \Lambda_i/\lambda_i = g_i(m_i)/(m_i \wedge \kappa_i x)$.

Summing up, the unique invariant NFMS is given by

$$z_i = g_i(m_i)/(m_i \wedge \kappa_i x) \text{ for } i \in \mathcal{I}_{\text{over}} \quad \text{and} \quad z_i = g_i(\kappa_i x)/(\kappa_i x) \text{ for } i \notin \mathcal{I}_{\text{over}},$$

where x is the unique solution to (27) and $\mathcal{I}_{\text{over}}$ is defined by (28).

7 Proof of Theorem 5

To prove **C**-tightness (that is, tightness with all weak limits being a.s. continuous) of $\{\overline{\mathcal{Z}}^r(\cdot)\}_{r \in \mathcal{R}}$, we check standard conditions (see e.g. [10]) of compact containment (Section 7.2) and oscillation control (Section 7.4). In Section 7.6, we check that fluid limits satisfy the fluid model equation (3).

To establish these main steps of the proof, we develop a number of auxiliary results. Section 7.1 contains a law of large numbers result for the load process. Section 7.3 proves that, for large r, $\overline{\mathcal{Z}}^r(\cdot)$ puts arbitrarily small mass to thin horizontal and vertical strips, which in particular implies

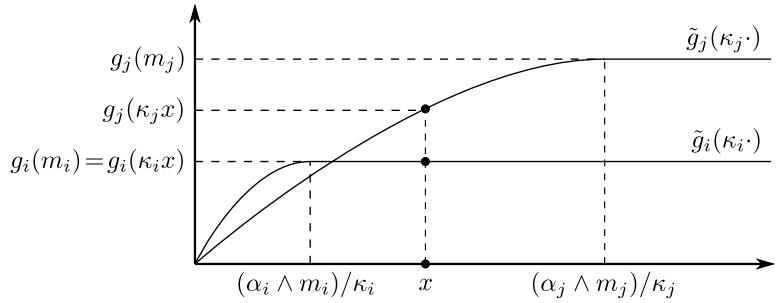

Figure 3: Graphs of the functions $\tilde{g}_i(\cdot)$; x is the unique solution to (27).

that fluid limits have both projections free of atoms. In Section 7.5, fluid limits are shown to be coordinate-wise bounded away from zero outside $t = 0$.

7.1 Load process

Introduce the measure valued load processes and their scaled versions: for all r, i and $t \geq s \geq 0$,

$$\begin{aligned} \mathcal{L}_i^r(t) &:= \sum\nolimits_{k=1}^{E_i^r(t)} \delta_{(B_{ik}^r, D_{ik}^r)}, & \overline{\mathcal{L}}_i^r(t) &:= \mathcal{L}_i^r(t)/r, \\ \mathcal{L}_i^r(s,t) &:= \mathcal{L}_i^r(t) - \mathcal{L}_i^r(s), & \overline{\mathcal{L}}_i^r(s,t) &:= \overline{\mathcal{L}}_i^r(t) - \overline{\mathcal{L}}_i^r(s). \end{aligned}$$

The following property is useful when proving other results of the section. Only minor adjustments in the proof of [13, Theorem 5.1] are needed to establish it.

Lemma 8. *By* (A.3) *and* (A.4)*, as* $r \to \infty$*,* $(\overline{\mathcal{L}}^r(\cdot), \langle \chi_1, \overline{\mathcal{L}}^r(\cdot)\rangle, \langle \chi_2, \overline{\mathcal{L}}^r(\cdot)\rangle) \Rightarrow (\eta(\cdot) * \theta, \rho(\cdot), \sigma(\cdot))$*, where* $\chi_1(x_1, x_2) := x_1$*,* $\chi_2(x_1, x_2) := x_2$*, and* $\eta(t) := t\eta$*,* $\rho(t) := t\rho$*,* $\sigma(t) := t\sigma$*.*

7.2 Compact containment

The property we prove here, together with the oscillation control result that follows in Section 7.4, implies tightness of the scaled state descriptors.

Lemma 9. *By* (A.3)–(A.5)*, for any* $T > 0$ *and* $\varepsilon > 0$*, there exists a compact set* $K \subset \mathbf{M}^I$ *such that*

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\overline{\mathcal{Z}}^r(t) \in K \text{ for all } t \in [0,T]\} \geq 1 - \varepsilon.$$

Proof. Fix T and ε. It suffices to show that, for each i, there exist a compact set $K_i \subset \mathbf{M}$ such that

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\overline{\mathcal{Z}}_i^r(t) \in K_i \text{ for all } t \in [0,T]\} \geq 1 - \varepsilon/I. \tag{29}$$

We use the following criterion (see e.g. [15, Theorem 15.7.5]).

Proposition 1. *A set* $\mathcal{M} \subset \mathbf{M}$ *is relatively compact if and only if* $\sup_{\xi\in\mathcal{M}} \xi(\mathbb{R}_+^2) < \infty$ *and* $\sup_{\xi\in\mathcal{M}} \xi(\mathbb{R}_+^2 \setminus [0,n]^2) \to 0$ *as* $n \to \infty$*.*

Note that

$$\overline{\mathcal{Z}}_i^r(t)(\mathbb{R}_+^2) = \overline{Z}_i^r(t) \leq \overline{Z}_i^r(0) + \overline{E}_i^r(T) = \overline{\mathcal{Z}}_i^r(0)(\mathbb{R}_+^2) + \overline{\mathcal{L}}_i^r(T)(\mathbb{R}_+^2). \tag{30}$$

Also note that, if the residual size (patience time) of a flow at time t exceeds n, then its initial size (patience time), must have exceeded n, too, which implies the following bound:

$$\overline{\mathcal{Z}}_i^r(t)(\mathbb{R}_+^2 \setminus [0,n]^2) \leq \overline{\mathcal{Z}}_i^r(0)(\mathbb{R}_+^2 \setminus [0,n]^2) + \overline{\mathcal{L}}_i^r(T)(\mathbb{R}_+^2 \setminus [0,n]^2). \tag{31}$$

The sequence $\{\overline{\mathcal{Z}}_i^r(0) + \overline{\mathcal{L}}_i^r(T)\}_{r\in\mathcal{R}}$ converges and hence in tight, i.e. there exists a compact set $K_i' \subset \mathbf{M}$ such that

$$\inf_{r\in\mathcal{R}} \mathbb{P}^r\{\overline{\mathcal{Z}}_i^r(0) + \overline{\mathcal{L}}_i^r(T) \in K_i'\} \geq 1 - \varepsilon/I. \tag{32}$$

Put

$$K_i'' := \{\xi \in \mathbf{M}\colon \text{for some } \xi' \in K_i',\ \xi(\mathbb{R}_+^2) \leq \xi'(\mathbb{R}_+^2) \text{ and } \xi(\mathbb{R}_+^2 \setminus [0,n]^2) \leq \xi'(\mathbb{R}_+^2 \setminus [0,n]^2),\ n \in \mathbb{N}\}.$$

Then the criterion of relative compactness for K_i'' follows from that for K_i', and (30)–(32) imply (29) with K_i taken as the closure of K_i''. □

7.3 Asymptotic regularity

This section contains three Lemmas. Lemmas 10 and 11 prove that neither initial nor newly arriving flows concentrate along horizontal and vertical lines. These two results are combined in Lemma 12 that implies the oscillation control result of the next section, and also is useful when deriving the limiting equations for the state descriptors in Section 7.6.

Recall from Section 5 that, for $b \geq a \geq 0$,

$$H_a^b = \mathbb{R}_+ \times [a,b], \quad V_a^b = [a,b] \times \mathbb{R}_+,$$

and introduce similar notations

$$H_a^\infty := \mathbb{R}_+ \times [a,\infty), \quad V_a^\infty := [a,\infty) \times \mathbb{R}_+.$$

Lemma 10. *By* (A.5) *and* (A.6)*, for any $\delta > 0$ and $\varepsilon > 0$, there exists an $a > 0$ such that*

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\sup_{x\in\mathbb{R}_+} \|\overline{\mathcal{Z}}^r(0)(H_x^{x+a})\| \vee \|\overline{\mathcal{Z}}^r(0)(V_x^{x+a})\| \leq \delta\} \geq 1 - \varepsilon.$$

Proof. Fix δ and ε. Since, for any $\xi \in \mathbf{M}^I$ and $a > 0$,

$$\sup_{x\in\mathbb{R}_+} \|\xi(H_x^{x+a})\| \vee \|\xi(V_x^{x+a})\| \leq 2 \sup_{n\in\mathbb{N}} \|\xi(H_{(n-1)a}^{na})\| \vee \|\xi(V_{(n-1)a}^{na})\|,$$

it suffices to find an a such that

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\overline{\mathcal{Z}}^r(0) \in \mathcal{M}_a\} \geq 1 - \varepsilon,$$

where $\mathcal{M}_a := \{\xi \in \mathbf{M}^I\colon \sup_{n\in\mathbb{N}} \|\xi(H_{(n-1)a}^{na})\| \vee \|\xi(V_{(n-1)a}^{na})\| < \delta/2\}$.

The set $\mathcal{M}_a$ is open because $\xi^k \xrightarrow{\mathrm{w}} \xi \in \mathcal{M}_a$ implies that $\xi^k \in \mathcal{M}_a$ for k large enough. Indeed, pick an $N \in \mathbb{N}$ such that $\|\xi(H_{Na}^\infty)\| \vee \|\xi(V_{Na}^\infty)\| < \delta/2$. Then, by the Portmanteau theorem,

$$\begin{aligned}
&\overline{\lim}_{k\to\infty} \sup_{n\in\mathbb{N}} \|\xi^k(H_{(n-1)a}^{na})\| \vee \|\xi^k(V_{(n-1)a}^{na})\| \\
\leq &\overline{\lim}_{k\to\infty} \max_{1\leq n\leq N} \|\xi^k(H_{(n-1)a}^{na})\| \vee \|\xi^k(V_{(n-1)a}^{na})\| \vee \|\xi^k(H_{Na}^\infty)\| \vee \|\xi^k(V_{Na}^\infty)\| \\
\leq &\max_{1\leq n\leq N} \|\xi(H_{(n-1)a}^{na})\| \vee \|\xi(V_{(n-1)a}^{na})\| \vee \|\xi(H_{Na}^\infty)\| \vee \|\xi(V_{Na}^\infty)\| < \delta/2.
\end{aligned}$$

By (A.6) and [13, Lemma A.1], there exists an a such that $\mathbb{P}\{\zeta^0 \in \mathcal{M}_a\} \geq 1 - \varepsilon$. Then, again by the Portmanteau theorem,

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\overline{\mathcal{Z}}^r(0) \in \mathcal{M}_a\} \geq \mathbb{P}\{\zeta^0 \in \mathcal{M}_a\} \geq 1 - \varepsilon.$$ □

Besides being used in the proof of the fluid limit theorem, the following result is also used when establishing convergence of the stationary distributions of the scaled state descriptors, see Section 5.

Lemma 11. *By* (A.3) *and* (A.4)*, for any $T > 0$, $\delta > 0$ and $\varepsilon > 0$, there exists an $a > 0$ such that*

$$\underline{\lim}_{r\to\infty}\mathbb{P}^r\underbrace{\{\sup_{t\in[0,T]}\sup_{x\in\mathbb{R}_+}\|\overline{\mathcal{Z}}^{r,\,new}(t)(H_x^{x+a})\|\vee\|\overline{\mathcal{Z}}^{r,\,new}(t)(V_x^{x+a})\|\le\delta\}}_{=:\,\Omega_*^r}\ge 1-\varepsilon.$$

Proof. Fix T, δ and ε. We first construct auxiliary events Ω_0^r such that $\underline{\lim}_{r\to\infty}\mathbb{P}^r\{\Omega_0^r\}\ge 1-\varepsilon$, and then show that $\Omega_*^r\supseteq\Omega_0^r$ for all r, which implies the theorem.

Definition of Ω_0^r. By Lemma 9, there exists a compact set $K\subset\mathbf{M}^I$ such that

$$\underline{\lim}_{r\to\infty}\mathbb{P}^r\underbrace{\{\overline{\mathcal{Z}}^r(t)\in K\text{ for all }t\in[0,T]\}}_{=:\,\Omega_1^r}\ge 1-\varepsilon,$$

and by Proposition 1, $M:=\sup_{\xi\in K}\|\xi(\mathbb{R}_+^2)\|<\infty$ and $\sup_{\xi\in K}\|\xi(\mathbb{R}_+^2)\setminus[0,L]^2\|\le\delta/4$ for a large enough L.

For each i, the rate function $\lambda_i(\cdot)$ is positive on $\{z\in\mathbb{R}_+^I\colon z_i>0\}$ and, by Lemma 1, it is continuous there. Hence,

$$\lambda_*:=\min_{1\le i\le I}\inf\{\lambda_i(z)\colon z_i\ge\delta/4,\ \|z\|\le M\}>0. \tag{33}$$

Put

$$\gamma:=\frac{\delta}{72\|\eta\|}\wedge T\quad\text{and}\quad a:=\frac{\gamma(\lambda_*\wedge 1)}{3}.$$

Also pick an N large enough so that

$$Na>L+(\|m\|\vee 1)T.$$

For $m,n\in\mathbb{N}$, define the sets

$$\begin{aligned}I_{m,n}&:=[(m-1)a,ma)\times[(n-1)a,na),\\ I^{m,n}&:=[(m-2)^+a,(m+1)a)\times[(n-2)^+a,(n+1)a),\end{aligned}$$

and pick functions $g_{m,n}\in\mathbf{C}_{\mathbb{R}_+^2\to[0,1]}$ such that

$$\mathbb{I}_{I_{m,n}}(\cdot)\le g_{m,n}(\cdot)\le\mathbb{I}_{I^{m,n}}(\cdot).$$

Since θ is a vector of probability measures,

$$\sum\nolimits_{m,n\in\mathbb{N}}\|\langle g_{m,n},\theta\rangle\|\le\|\sum\nolimits_{m,n\in\mathbb{N}}\theta(I^{m,n})\|\le 9. \tag{34}$$

By Lemma 8 and the continuous mapping theorem, for all $m,n\in\mathbb{N}$, $\langle g_{m,n},\overline{\mathcal{L}}^r(\cdot)\rangle\Rightarrow\eta(\cdot)\langle g_{m,n},\theta\rangle$ as $r\to\infty$. Since the limits are deterministic, we have convergence in probability. Since the limits are continuous, we have uniform convergence on compact sets. Hence,

$$\lim_{r\to\infty}\mathbb{P}^r\underbrace{\{\max_{1\le m,n\le N}\sup_{t\in[0,T]}\|\langle g_{m,n},\overline{\mathcal{L}}^r(t)\rangle-t\eta*\langle g_{m,n},\theta\rangle\|\le\delta/(16N^2)\}}_{=:\,\Omega_2^r}=1.$$

Similarly, by (A.3),

$$\lim_{r\to\infty}\mathbb{P}^r\underbrace{\{\sup_{t\in[0,T]}\|\overline{E}^r(t)-t\eta\|\le\delta/16\}}_{=:\,\Omega_3^r}=1.$$

For all r, put

$$\Omega_0^r:=\Omega_1^r\cap\Omega_2^r\cap\Omega_3^r,$$

then $\underline{\lim}_{r\to\infty}\mathbb{P}^r\{\Omega_0^r\}\ge 1-\varepsilon$, and it is left to show that $\Omega_0^r\subseteq\Omega_*^r$.

Proof of $\Omega_0^r \subseteq \Omega_*^r$. Fix $r \in \mathcal{R}$, $t \in [0,T]$, $x \in \mathbb{R}_+$ and i. Also fix an outcome $\omega \in \Omega_0^r$. All random objects in the rest of the proof will be evaluated at this ω. We have to check that

$$\overline{\mathcal{Z}}_i^{r,\,\mathrm{new}}(t)(H_x^{x+a}) \le \delta, \tag{35a}$$

$$\overline{\mathcal{Z}}_i^{r,\,\mathrm{new}}(t)(V_x^{x+a}) \le \delta. \tag{35b}$$

We will show (35a), (35b) follows similarly.

Define the random time $\tau := \sup\{s \le t\colon \overline{Z}_i^{r,\,\mathrm{new}}(s) < \delta/4\}$ (supremum over the empty set equals 0 by convention). Although in general τ is not a continuity point for $\overline{Z}_i^{r,\,\mathrm{new}}(\cdot)$, we still can estimate $\overline{Z}_i^{r,\,\mathrm{new}}(\tau)$:

$$\overline{Z}_i^{r,\,\mathrm{new}}(\tau) \le \delta/2. \tag{36}$$

Indeed, if $\tau = 0$, then $\overline{Z}_i^{r,\,\mathrm{new}}(\tau) = 0$, and (36) holds. If $\tau > 0$, pick a $\tau' \in [(\tau-\gamma)^+, \tau]$ such that $\overline{Z}_i^{r,\,\mathrm{new}}(\tau') < \delta/4$. Then, by the definition of Ω_3^r,

$$\overline{Z}_i^{r,\,\mathrm{new}}(\tau) \le \overline{Z}_i^{r,\,\mathrm{new}}(\tau') + (\overline{E}_i^r(\tau) - \overline{E}_i^r(\tau')) \le \delta/4 + \eta_i(\tau-\tau') + \delta/8 \le \|\eta\|\gamma + 3\delta/8,$$

and (36) holds by the choice of γ.

Now, if $\tau = t$, then (36) implies (35a), and the proof is finished. Assume that $\tau < t$. Then, by the choice of L and (36),

$$\begin{aligned}\overline{\mathcal{Z}}_i^{r,\,\mathrm{new}}(t)(H_x^{x+a}) &\le \overline{\mathcal{Z}}_i^{r,\,\mathrm{new}}(t)(H_x^{x+a} \cap [0,L]^2) + \delta/4\\ &\le \underbrace{\overline{Z}_i^{r,\,\mathrm{new}}(\tau)}_{\le \delta/2} + \frac{1}{r}\sum_{E_i^r(\tau)+1}^{E_i^r(t)} \underbrace{\mathbb{I}_{H_x^{x+a}\cap[0,L]^2}(B_{ik}^r - S_i(\overline{Z}^r, U_{ik}^r, t), D_{ik}^r - (t-U_{ik}^r))}_{=:\,s_k} + \delta/4\end{aligned}$$

and in order to have (35a), it suffices to show that

$$\Sigma := \frac{1}{r}\sum_{E_i^r(\tau)+1}^{E_i^r(t)} s_k = \sum_{m,n\in\mathbb{N}} \underbrace{\frac{1}{r}\sum_{E_i^r(\tau)+1}^{E_i^r(t)} s_k \mathbb{I}_{I_{m,n}}(B_{ik}^r, D_{ik}^r)}_{=:\,\Sigma_{m,n}} \le \delta/4. \tag{37}$$

First note that

$$\Sigma_{m,n} = 0 \quad \text{if} \quad m > N \text{ or } n > N. \tag{38}$$

Indeed, consider a flow on route i that arrived at $U_{ik}^r \in (\tau, t]$ with $(B_{ik}^r, D_{ik}^r) \in I_{m,n}$. If $m > N$, then $B_{ik}^r > L + \|m\|T$ by the choice of N, $B_{ik}^r - S_i(\overline{Z}^r, U_{ik}^r, t) > L$ by the rate constraints, and $s_k = 0$. If $n > N$, then $D_{ik}^r > L + T$ by the choice of N, $D_{ik}^r - (t - U_{ik}^r) > L$ and again $s_k = 0$.

Now we estimate $\Sigma_{m,n}$ for $1 \le m, n \le N$. Fix m, n. Consider two flows $k < l$ such that $U_{ik}^r, U_{il}^r \in (\tau, t]$ and $(B_{ik}^r, D_{ik}^r), (B_{il}^r, D_{il}^r) \in I_{m,n}$. In $(\tau, t]$, $\overline{Z}_i^r(\cdot) \ge \overline{Z}_i^{r,\,\mathrm{new}}(\cdot) \ge \varepsilon/4$ and $\|\overline{Z}^r(\cdot)\| \le M$, and then (33) implies that

$$\inf_{s\in(\tau,t]} \lambda_i(\overline{Z}^r(s)) \ge \lambda_*.$$

If $U_{il}^r - U_{ik}^r \ge \gamma$, then

$$\begin{aligned}(B_{il}^r - S_i(\overline{Z}^r, U_{il}^r, t)) - (B_{ik}^r - S_i(\overline{Z}^r, U_{ik}^r, t)) &\ge \overbrace{\gamma\lambda_*}^{\ge 3a} - \overbrace{(B_{ik}^r - B_{il}^r)}^{\le a} \ge 2a,\\ (D_{il}^r - (t-U_{il}^r)) - (D_{ik}^r - (t-U_{ik}^r)) &\ge \underbrace{\gamma}_{\ge 3a} - \underbrace{(D_{ik}^r - D_{il}^r)}_{\le a} \ge 2a,\end{aligned}$$

and hence at most one of s_k and s_l is non-zero. This implies that all arrivals to route i during $(\tau, t]$ that correspond to non-zero summands in $\Sigma_{m,n}$ occur actually during a smaller interval $(t_{m,n}, t_{m,n} + \gamma] \subseteq (\tau, t]$. Then, by the definition of Ω_2^r,

$$\begin{aligned}\Sigma_{m,n} &\le \frac{1}{r}\sum_{k=E_i^r(t_{m,n})+1}^{E_i^r(t_{m,n}+\gamma)} \mathbb{I}_{I_{m,n}}(B_{ik}^r, D_{ik}^r) \le \sup_{s\in[0,T-\gamma]} \frac{1}{r}\sum_{k=E_i^r(s+1)}^{E_i^r(s+\gamma)} g_{m,n}(B_{ik}^r, D_{ik}^r)\\ &= \sup_{s\in[0,T-\gamma]}(\langle g_{m,n}, \overline{\mathcal{L}}_i^r(s+\gamma)\rangle - \langle g_{m,n}, \overline{\mathcal{L}}_i^r(s)\rangle) \le \gamma\eta_i\langle g_{m,n}, \theta_i\rangle + \delta/(8N^2).\end{aligned}$$

We plug the last inequality and (38) into $\Sigma = \sum_{m,n\in\mathbb{N}} \Sigma_{m,n}$, then (37) follows by (34) and the choice of γ. □

The previous two lemmas are summed up into the following result.

Lemma 12. *By* (A.3)–(A.6)*, for any $T > 0$, $\delta > 0$ and $\varepsilon > 0$, there exists an $a > 0$ such that*

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \|\overline{\mathcal{Z}}^r(t)(H_x^{x+a})\| \vee \|\overline{\mathcal{Z}}^r(t)(V_x^{x+a})\| \le \delta\} \ge 1-\varepsilon.$$

Proof. Note that

$$\begin{aligned}&\sup_{x\in\mathbb{R}_+} \|\overline{\mathcal{Z}}^{r,\,\mathrm{init}}(t)(H_x^{x+a})\| \vee \|\overline{\mathcal{Z}}^{r,\,\mathrm{init}}(t)(V_x^{x+a})\| \\ \le &\sup_{x\in\mathbb{R}_+} \|\overline{\mathcal{Z}}^{r}(0)(H_x^{x+a})\| \vee \|\overline{\mathcal{Z}}^{r}(0)(V_x^{x+a})\|.\end{aligned}$$

Indeed,

$$\overline{\mathcal{Z}}_i^{r,\,\mathrm{init}}(t)(H_x^{x+a}) \le \overline{\mathcal{Z}}_i^r(0)(H_{x+t}^{x+a+t}) \quad \text{and} \quad \overline{\mathcal{Z}}_i^{r,\,\mathrm{init}}(t)(V_x^{x+a}) \le \overline{\mathcal{Z}}_i^r(0)(V_{x+S_i(\overline{Z}^r,0,t)}^{x+a+S_i(\overline{Z}^r,0,t)}).$$

Then the lemma follows by $\overline{\mathcal{Z}}^r(\cdot) = (\overline{\mathcal{Z}}^{r,\,\mathrm{init}} + \overline{\mathcal{Z}}^{r,\,\mathrm{new}})(\cdot)$ and Lemmas 10 and 11. □

7.4 Oscillation control

Here we establish the second key ingredient of tightness of the scaled state descriptors, the first one is proven in Section 7.2.

Lemma 13. *By* (A.3)–(A.6)*, for any $T > 0$, $\delta > 0$ and $\varepsilon > 0$, there exists an $h > 0$ such that*

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\underbrace{\omega(\overline{\mathcal{Z}}^r, h, T) \le \delta}_{=:\,\Omega_*^r}\} \ge 1-\varepsilon,$$

where $\omega(\overline{\mathcal{Z}}^r, h, T) := \sup\{d_I(\overline{\mathcal{Z}}^r(s), \overline{\mathcal{Z}}^r(t))\colon s,t\in[0,T],\ |s-t|<h\}$.

Proof. Fix T, δ and ε. By (A.3),

$$\lim_{r\to\infty} \mathbb{P}^r\{\underbrace{\sup_{t\in[0,T]} \|\overline{E}^r(t) - t\eta\| \le \delta/4}_{=:\,\Omega_1^r}\} = 1.$$

By Lemma 12, there exists an $a > 0$ such that

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\underbrace{\sup_{t\in[0,T]} \|\overline{\mathcal{Z}}^r(t)(H_0^a \cup V_0^a)\| \le \delta}_{=:\,\Omega_2^r}\} \ge 1-\varepsilon.$$

Pick an h such that $h(\|m\| \vee 1) \le \delta \wedge a$ and $h\|\eta\| \le \delta/2$. We now show that, for all $r \in \mathcal{R}$, $\Omega_*^r \supseteq \Omega_1^r \cap \Omega_2^r$, then the lemma follows.

Fix $r \in \mathcal{R}$, i and $s,t \in [0,T]$ such that $s < t$, $t - s < h$. Also fix an outcome $\omega \in \Omega_1^r \cap \Omega_2^r$. All random objects in the rest of the proof will be evaluated at this ω. We have to check that, for any non-empty closed Borel subset $B \subseteq \mathbb{R}_+^2$,

$$\overline{\mathcal{Z}}_i^r(s)(B) \le \overline{\mathcal{Z}}_i^r(t)(B^\delta) + \delta, \tag{39a}$$

$$\overline{\mathcal{Z}}_i^r(t)(B) \le \overline{\mathcal{Z}}_i^r(s)(B^\delta) + \delta. \tag{39b}$$

First we check (39a). Note that it suffices to show

$$\overline{\mathcal{Z}}_i^r(s)(B) \le \overline{\mathcal{Z}}_i^r(\tau)(B^\delta) + \delta, \tag{40}$$

where $\tau := \inf\{u \in [s,t]\colon \overline{Z}_i^r(u) = 0\}$ and infimum over the empty set equals t by definition. Indeed, if $\tau = t$, then (40) implies (39a). If $\tau < t$, then by the right-continuity of $\overline{Z}_i^r(\cdot)$, $\overline{\mathcal{Z}}_i^r(\tau)(B^\delta) = \overline{Z}_i^r(\tau) = 0$, and again (40) implies (39a).

Now prove (40). If $\tau = s$, then (40) holds. Assume that $\tau > s$. By the defintion of Ω_2^r,

$$\overline{\mathcal{Z}}_i^r(s)(B) \le \overline{\mathcal{Z}}_i^r(s)(B \cap [a,\infty)^2) + \delta. \tag{41}$$

Since $S_i(\overline{Z}^r, s, \tau) < \|m\|h \le \delta \wedge a$ and $\tau - s < h \le \delta \wedge a$,

$$\overline{\mathcal{Z}}_i^r(s)(B \cap [a,\infty)^2) \le \overline{\mathcal{Z}}_i^r(\tau)(B^\delta),$$

which together with (41) implies (40).

It is left to check (39b). Since $S_i(\overline{Z}^r, s, \tau) < \|m\|h \le \delta$ and $\tau - s < h \le \delta$,

$$\overline{\mathcal{Z}}_i^r(t)(B) \le \overline{\mathcal{Z}}_i^r(s)(B^\delta) + (\overline{E}_i^r(t) - \overline{E}_i^r(s)),$$

and (39b) follows by the definition of Ω_1^r. □

7.5 Fluid limits are bounded away from zero

Rate constraints provide infinite-server-queue lower bounds for bandwidth-sharing networks. First we show that properly scaled infinite server queues are bounded away from zero, and then the same follows for bandwidth-sharing networks with rate constraints.

Consider a sequence of infinite server queues marked by $r \in \mathcal{R}$. At $t = 0$, the queues are empty. To the r-th queue, customers arrive according to a counting process $A^r(\cdot)$ and have i.i.d. service times $\{B_k^r\}_{k\in\mathbb{N}}$ distributed as B^r. Let $\overline{A}^r(\cdot) := A^r(\cdot)/r \Rightarrow \alpha(\cdot)$, where $\alpha(t) := t\alpha$ and $\alpha > 0$. Also let $B^r \Rightarrow B$, where $\mathbb{P}\{B > 0\} > 0$. Denote by $Q^r(\cdot)$ the population process of the r-th queue and put $\overline{Q}^r(\cdot) := Q^r(\cdot)/r$.

Lemma 14. *For any $\delta > 0$, there exists a $C(\delta) > 0$ such that, for any $\Delta > \delta$,*

$$\mathbb{P}^r\{\inf_{\delta\le t\le\Delta} \overline{Q}^r(t) \ge C(\delta)\} \to 1 \quad \text{as } r \to \infty.$$

Proof. Let us first explain the result heuristically. Consider the arrivals with long service times, i.e. exceeding a $b > 0$. During $(0, b/2]$, there are $r\alpha\mathbb{P}\{B > b\}b/2$ such arrivals to the r-th queue. They will leave the queue after $t = b$, and hence, in $(b/2, b]$, the scaled queue length $\overline{Q}^r(\cdot)$ is bounded from below by $\alpha\mathbb{P}\{B > b\}b/2$. Similarly, $\overline{Q}^r(\cdot) \ge \alpha\mathbb{P}\{B > b\}b/2$ in any interval $((n-1)b/2, nb/2]$, $n \in \mathbb{N}$.

We now proceed more formally. Pick an $b \in (0, \delta)$ such that b is a continuity point for the distribution of B, and

$$p := \mathbb{P}\{B \ge b\} > 0.$$

Then, as $r \to \infty$,

$$p_r := \mathbb{P}^r\{B^r \ge b\} \to p.$$

Partition $(0, \Delta]$ into subintervals of length $b/2$,

$$(0,\Delta] \subseteq \bigcup\nolimits_{1\le n\le N(\Delta)} ((n-1)b/2, nb/2].$$

Denote by $\overline{A}_n^r$ the scaled number of arrivals during $((n-1)b/2, nb/2]$, and by $\overline{A}_n^r(b)$ the scaled number of arrivals during $((n-1)b/2, nb/2]$ with service times at least b,

$$\overline{A}_n^r := \overline{A}^r(nb/2) - \overline{A}^r((n-1)b/2),$$
$$\overline{A}_n^r(b) := \frac{1}{r}\sum\nolimits_{k=A^r((n-1)b/2)+1}^{A^r(nb/2)} \mathbb{I}_{\{B_k^r \ge b\}}.$$

By $\overline{A}^r \Rightarrow \alpha(\cdot)$ and $p_r \to p$ as $r \to \infty$,

$$(\overline{A}^r_1, \ldots, \overline{A}^r_{N(\Delta)}) \Rightarrow (\alpha b/2, \ldots, \alpha b/2),$$
$$(\overline{A}^r_1(b), \ldots, \overline{A}^r_{N(\Delta)}(b)) \Rightarrow (\alpha p b/2, \ldots, \alpha p b/2).$$

Pick a $C(\delta) < \alpha p b/2$, then

$$\begin{aligned}
&\mathbb{P}^r\{\inf_{\delta \le t \le \Delta} \overline{Q}^r(t) \ge C(\delta)\} \\
&\ge \mathbb{P}^r\{\inf_{t \in ((n-1)b/2, nb/2]} \overline{Q}^r(t) \ge C(\delta),\ n = 2, \ldots, N(\Delta)\} \\
&\ge \mathbb{P}^r\{\overline{A}^r_n(b) \ge C(\delta),\ n = 1, \ldots, N(\Delta) - 1\} \to 1 \quad \text{as } r \to \infty.
\end{aligned}$$

□

We can now prove easily that all fluid limits are bounded away from zero outside $t = 0$.

Lemma 15. *For any $\delta > 0$, there exists a $C(\delta) > 0$ such that, for any fluid limit $(\mathcal{Z}, Z)(\cdot)$,*

$$\text{a.s.} \quad \inf_{t \ge \delta} \min_{1 \le i \le I} Z_i(t) \ge C(\delta).$$

Proof. Consider a flow k on route i in the r-th network. By the rate constraints, this flow will stay in the network at least for $B^r_{ik}/m_i \wedge D^r_{ik}$ since its arrival. Hence, the route i population process $Z^r_i(\cdot)$ is bounded from below by the length $Q^r_i(\cdot)$ of the infinite server queue with arrivals $E^r_i(\cdot)$ and i.i.d. service times $\{B^r_{ik}/m_i \wedge D^r_{ik}\}_{k \in \mathbb{N}}$. Assume that $Q^r_i(0) = 0$ and put $\overline{Q}^r_i(\cdot) = Q^r_i(\cdot)/r$. Then, by Lemma 14, for any $\delta > 0$ there exists a $C(\delta) > 0$ such that, for any $\Delta > \delta$,

$$\mathbb{P}^r\{\inf_{t \in [\delta, \Delta]} \min_{1 \le i \le I} \overline{Z}^r_i(t) \ge C(\delta)\} \ge \mathbb{P}^r\{\inf_{t \in [\delta, \Delta]} \min_{1 \le i \le I} \overline{Q}^r_i(t) \ge C(\delta)\} \to 1.$$

Now consider a fluid limit $(\mathcal{Z}, Z)(\cdot)$ along a subsequence $\{(\overline{\mathcal{Z}}^q, \overline{Z}^q)(\cdot)\}_{q \in \mathcal{Q}}$. For any compact set $K \subset \mathbb{R}_+$, the mapping $\varphi_K : \mathbf{D}_{\mathbb{R}_+ \to \mathbb{R}} \to \mathbb{R}$, $\varphi_K(x) := \inf_{t \in K} \min_{q \le i \le I} x(t)$ is continuous at continuous $x(\cdot)$. Hence, $\varphi_{[\delta,\Delta]}(\overline{Z}^q) \Rightarrow \varphi_{[\delta,\Delta]}(Z)$ and, by the Portmanteau theorem,

$$\mathbb{P}\{\varphi_{[\delta,\Delta]}(Z) \ge C(\delta)\} \ge \overline{\lim}_{q \to \infty} \mathbb{P}^q\{\varphi_{[\delta,\Delta]}(\overline{Z}^q) \ge C(\delta)\} = 1,$$

where $\Delta > \delta$ is arbitrary. Then the lemma follows.

Note also that the constant $C(\delta)$ does not depend on a particular fluid limit $(\mathcal{Z}, Z)(\cdot)$. □

7.6 Fluid limits as fluid model solutions

Here we show that fluid limits a.s. satisfy the fluid model equation (3).

Let $(\mathcal{Z}, Z)(\cdot)$ be a fluid limit along a subsequence $\{(\overline{\mathcal{Z}}^q, \overline{Z}^q)(\cdot)\}_{q \in \mathcal{Q}}$. Lemma 12 implies that (*cf.* the proof of [11, Lemma 6.2])

$$\text{a.s.} \quad \mathcal{Z}_i(t)(\partial_A) = 0 \quad \text{for all } t \ge 0, \text{ all } i \text{ and } A \in \mathcal{C}, \tag{42}$$

where ∂_A denotes the boundary of A. Then, when proving (3) for $(\mathcal{Z}, Z)(\cdot)$, it suffices to consider sets A from

$$\mathcal{C}^+ := \{[x, \infty) \times [y, \infty) \colon x \wedge y > 0\}.$$

It also suffices to consider t from a finite interval $[0, T]$.

The rest of the proof splits into two parts. First we derive dynamic equations for the prelimiting processes $(\overline{\mathcal{Z}}^q, \overline{Z}^q)(\cdot)$, and then show that these equations converge to (3).

Prelimiting equations Fix $q \in \mathcal{Q}$, i, $t \le T$ and $A \in \mathcal{C}^+$. Fix also an outcome $\omega \in \Omega^q$. In what follows up to equation (45), all random elements are evaluated at this ω. We have

$$\begin{aligned}\overline{\mathcal{Z}}_i^q(t)(A) =& \overline{\mathcal{Z}}_i^q(0)(A + (S_i(\overline{Z}^q, 0, t), t)) \\ &+ \underbrace{\frac{1}{q}\sum_{k=1}^{E_i^q(t)} \overbrace{\mathbb{I}_A(B_{ik}^q - S_i(\overline{Z}^q, U_{ik}^q, t), D_{ik}^q - (t - U_{ik}^q))}^{=:\, s_k}}_{=:\, \Sigma}.\end{aligned} \tag{43}$$

Fix a partition partition $0 < t_0 < t_1 < ... < t_N = t$, then

$$\Sigma = \frac{1}{q}\sum_{k=1}^{E_i^q(t_0)} s_k + \frac{1}{q}\sum_{j=0}^{N-1}\sum_{k=E_i^q(t_j)+1}^{E_i^q(t_{j+1})} s_k.$$

Suppose that a function $y(\cdot)$ is non-increasing in $[t_0, t]$ and that, for some δ,

$$\sup_{s\in[t_0,t]} |S_i(\overline{Z}^q, s, t) - y(s)| \le \delta.$$

Now we can estimate Σ. If $U_{ik}^q \in (t_j, t_{j+1}]$, then

$$\begin{aligned}B_{ik}^q - (y(t_j) + \delta) \le B_{ik}^q - S(\overline{Z}^q, U_{ik}^q, t) \le B_{ik}^q - (y(t_{j+1}) - \delta), \\ D_{ik}^q - (t - t_j)) \le D_{ik}^q - (t - U_{ik}^q) \le D_{ik}^q - (t - t_{j+1}),\end{aligned}$$

and

$$\begin{aligned}\Sigma &\ge \sum_{j=0}^{N-1} \frac{1}{q}\sum_{k=E_i^q(t_j)+1}^{E_i^q(t_{j+1})} \mathbb{I}_A(B_{ik}^q - (y(t_j) + \delta), D_{ik}^q - (t - t_j)), \\ \Sigma &\le \overline{E}_i^q(t_0) + \sum_{j=0}^{N-1} \frac{1}{q}\sum_{k=E_i^q(t_j)+1}^{E_i^q(t_{j+1})} \mathbb{I}_A(B_{ik}^q - (y(t_{j+1}) - \delta), D_{ik}^q - (t - t_{j+1})),\end{aligned}$$

which can be rewritten as

$$\begin{aligned}\Sigma &\ge \sum_{j=0}^{N-1} \overline{\mathcal{L}}_i^q(t_j, t_{j+1})(A + (y(t_j) + \delta, t - t_j)) \\ \Sigma &\le \overline{E}_i^q(t_0) + \sum_{j=0}^{N-1} \overline{\mathcal{L}}_i^q(t_j, t_{j+1})(A + (y(t_{j+1}) - \delta, t - t_{j+1})).\end{aligned} \tag{44}$$

Put

$$X^q := \sup_{A\in\mathcal{C}} \sup_{0\le s\le t\le T} \|(\overline{\mathcal{L}}^q(s,t)(A) - (t-s)\eta * \theta^q(A)\|,$$

then, by (44) and (43),

$$\begin{aligned}&\sum_{j=0}^{N-1}\Big(\eta_i(t_{j+1} - t_j)\theta_i^q(A + (y(t_j) + \delta, t - t_j)) - X^q\Big) \\ &\le \overline{\mathcal{Z}}_i^q(t)(A) - \overline{\mathcal{Z}}_i^q(0)(A + (S_i(\overline{Z}^q, 0, t), t)) \\ &\le \eta_i t_0 + X^q + \sum_{j=0}^{N-1}\Big(\eta_i(t_{j+1} - t_j)\theta_i^q(A + (y(t_{j+1}) - \delta, t - t_{j+1}) + X^q\Big).\end{aligned} \tag{45}$$

To summarize, we have shown that, for all $q \in \mathcal{Q}$ and $\omega \in \Omega^q$,

$$(\overline{\mathcal{Z}}^q(\cdot), X^q) \in \mathcal{A}^q, \tag{46}$$

where $\mathcal{A}^q \subset \mathbf{D}_{\mathbb{R}_+\to\mathbf{M}^I} \times \mathbb{R}_+$ is the set of pairs $(\zeta(\cdot), x)$ such that, for any set $A \in \mathcal{C}^+$, any partition $0 < t_0 < t_1 < \ldots < t_N = t \le T$ and any function $y(\cdot)$ that is non-increasing in $[t_0, t]$ and that satisfies $\sup_{s\in[t_0,t]} |S_i(\langle 1, \zeta\rangle, s, t) - y(s)| \le \delta$ for some i and δ,

$$\begin{aligned}&\sum_{j=0}^{N-1}\Big(\eta_i(t_{j+1} - t_j)\theta_i^q(A + (y(t_j) + \delta, t - t_j)) - x\Big) \\ &\le \zeta(t)(A) - \zeta_i(0)(A + (S_i(\langle 1, \zeta\rangle, 0, t), t)) \\ &\le \eta_i t_0 + x + \sum_{j=0}^{N-1}\Big(\eta_i(t_{j+1} - t_j)\theta_i^q(A + (y(t_{j+1}) - \delta, t - t_{j+1}) + x\Big).\end{aligned}$$

Limiting equations By (A.3) *and* (A.4) (*cf.* the proof of [11, Lemma 5.1]),

$$X_q \Rightarrow 0 \quad \text{as } q \to \infty.$$

Since the limit of X_q is deterministic, then the joint convergence $(\overline{\mathcal{Z}}^q(\cdot), X^q) \Rightarrow (\mathcal{Z}(\cdot), 0)$ holds. By the Skorokhod representation theorem, there exist random elements $\{\tilde{\mathcal{Z}}^q(\cdot)\}_{q\in\mathcal{Q}}$, $\tilde{\mathcal{Z}}(\cdot)$ and $\{\tilde{X}^q\}_{q\in\mathcal{Q}}$ defined on a common probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ such that $(\tilde{\mathcal{Z}}^q(\cdot), \tilde{X}^q) \stackrel{\mathrm{d}}{=} (\overline{\mathcal{Z}}^q(\cdot), X^q)$, $q \in \mathcal{Q}$, and $\tilde{\mathcal{Z}}(\cdot) \stackrel{\mathrm{d}}{=} \mathcal{Z}(\cdot)$, and

$$\text{a.s.} \quad (\tilde{\mathcal{Z}}^q(\cdot), \tilde{X}^q) \to (\tilde{\mathcal{Z}}(\cdot), 0) \quad \text{as } q \to \infty. \tag{47}$$

Introduce also the total mass processes $\tilde{Z}^q(\cdot) := \langle 1, \tilde{\mathcal{Z}}^q(\cdot)\rangle$, $q \in \mathcal{Q}$, and $\tilde{Z}(\cdot) := \langle 1, \tilde{\mathcal{Z}}(\cdot)\rangle$. By Lemma 15, (42) and (46),

$$\text{a.s.} \quad \tilde{Z}_i(t) > 0 \quad \text{for all } t > 0 \text{ and all } i, \tag{48a}$$

$$\text{a.s.} \quad \tilde{\mathcal{Z}}_i(t)(\partial_A) = 0 \quad \text{for all } t \geq 0, \text{ all } i \text{ and } A \in \mathcal{C}, \tag{48b}$$

$$\text{a.s.} \quad (\tilde{\mathcal{Z}}^q(\cdot), \tilde{X}^q) \in \mathcal{A}^q \quad \text{for all } q \in \mathcal{Q}. \tag{48c}$$

Denote by $\tilde{\Omega}_*$ the set of outcomes $w \in \tilde{\Omega}$ for which (47) and (48) hold. We will show that, for all $\omega \in \tilde{\Omega}_*$, all i, $t \in [0, T]$ and $A \in \mathcal{C}^+$,

$$\begin{aligned} \tilde{\mathcal{Z}}_i(t)(A) =& \tilde{\mathcal{Z}}_i(0)(A + (S_i(\tilde{Z}, 0, t), t)) \\ &+ \eta_i \int_0^t \theta_i(A + (S_i(\tilde{Z}, s, t), t - s))ds, \end{aligned} \tag{49}$$

and that will complete the proof of Theorem 5.

Fix $t \in [0, T]$, i and $A \in \mathcal{C}^+$. Also fix an outcome $\omega \in \tilde{\Omega}_*$. All random elements in the rest of the proof are evaluated at this ω.

By (47) and (48b),

$$\tilde{\mathcal{Z}}_i^q(t)(A) \to \tilde{\mathcal{Z}}_i(t)(A) \quad \text{as } q \to \infty. \tag{50}$$

By (48a), the rate constraints and the dominated convergence theorem,

$$S_i(\tilde{Z}^q, s, t) \to S_i(\tilde{Z}, s, t) \quad \text{for all } s \in [0, t] \quad \text{as } q \to \infty, \tag{51}$$

which in particular implies that

$$\tilde{\mathcal{Z}}_i^q(0)(A + (S_i(\tilde{Z}^q, 0, t), t)) \to \tilde{\mathcal{Z}}_i(0)(A + (S_i(\tilde{Z}, 0, t), t)) \quad \text{as } q \to \infty. \tag{52}$$

Fix $t_0 \in (0, t)$ and $\delta > 0$. By (48a), the function $S_i(\tilde{Z}, \cdot, t)$ is continuous in $[t_0, t]$, and the functions $S_i(\tilde{Z}^q, \cdot, t)$ are monotone. Then the point-wise convergence (51) implies uniform convergence in $[t_0, t]$, and for q large enough,

$$\sup_{s\in[t_0,t]} |S_i(\tilde{Z}^q, s, t) - S_i(\tilde{Z}, s, t)| \leq \delta. \tag{53}$$

Now fix a partition $t_0 < t_1 < \ldots < t_N = t$. The bound (53) and (48c) imply that (in the definition of $\mathcal{A}^q$ we take $y(\cdot) = S_i(\tilde{Z}, \cdot, t)$)

$$\begin{aligned} &\sum\nolimits_{j=0}^{N-1} \Big(\eta_i(t_{j+1} - t_j)\theta_i^q(A + (S_i(\tilde{Z}, t_j, t) + \delta, t - t_j)) - \tilde{X}^q\Big) \\ &\leq \tilde{\mathcal{Z}}_i^q(t)(A) - \tilde{\mathcal{Z}}_i^q(A + (S_i(\tilde{Z}^q, 0, t), t)) \\ &\leq \eta_i t_0 + \tilde{X}^q + \sum\nolimits_{j=0}^{N-1} \Big(\eta_i(t_{j+1} - t_j)\theta_i^q(A + (S_i(\tilde{Z}, t_{j+1}, t) - \delta, t - t_{j+1}) + \tilde{X}^q\Big). \end{aligned} \tag{54}$$

Since $\theta_i(\cdot \times \mathbb{R}_+)$ and $\theta_i(\mathbb{R}_+ \times \cdot)$ are probability measures, the set of $B \in \mathcal{C}$ for which $\theta_i(\partial_B) > 0$ is at most countable. By (48), $S_i(\tilde{Z}, \cdot, t)$ is strictly monotone in $[t_0, t]$. Hence, the set $\mathcal{D}$ of $s \in [t_0, t]$ for which $\theta_i(\partial_{A+(S_i(\tilde{Z},s,t)+\delta,t-s)}) > 0$ or $\theta_i(\partial_{A+(S_i(\tilde{Z},s,t)-\delta,t-s)}) > 0$ is at most countable, too. In (54), let $q \to \infty$ assuming that the partition contains no points from $\mathcal{D}$. Then, by (47), (50)

and (52),

$$\begin{aligned}
&\sum_{j=0}^{N-1} \eta_i(t_{j+1}-t_j)\theta_i(A+(S_i(\tilde{Z},t_j,t)+\delta,t-t_j)) \\
&\le \tilde{\mathcal{Z}}_i(t)(A)-\tilde{\mathcal{Z}}_i(0)(A+(S_i(\tilde{Z},0,t),t)) \\
&\le \eta_i t_0+\sum_{j=0}^{N-1} \eta_i(t_{j+1}-t_j)\theta_i(A+(S_i(\tilde{Z},t_{j+1},t)-\delta,t-t_{j+1}).
\end{aligned} \tag{55}$$

Now, in (55), let the diameter of the partition go to 0 keeping t_0 fixed. Then

$$\begin{aligned}
&\eta_i \int_{t_0}^{t} \theta_i(A+(S_i(\tilde{Z},s,t)+\delta,t-s))ds \\
&\le \tilde{\mathcal{Z}}_i(t)(A)-\tilde{\mathcal{Z}}_i(0)(A+(S_i(\tilde{Z},0,t),t)) \\
&\le \eta_i t_0+\eta_i \int_{t_0}^{t} \theta_i(A+(S_i(\tilde{Z},s,t)-\delta,t-s)ds.
\end{aligned}$$

Finally, in the last inequality, let $\delta \to 0$ (recall (48b)) and $t_0 \to 0$, then (49) follows.

8 Proof of Theorem 6

By the discussion following Theorem 6 and Lemma 2, it is left to show tightness of the scaled stationary distributions. It suffices to show coordinate-wise tightness, so fix i. By [14, Theorem 2.1] and [15, Theorem 15.7.5], the sequence $\{\overline{\mathcal{Y}}_i^r, \overline{Y}_i^r\}_{r\in\mathcal{R}}$ is tight if

$$\sup_{r\in\mathcal{R}} \mathbb{E}^r \overline{Y}_i^r < \infty, \tag{56a}$$

$$\lim_{n\to\infty} \mathbb{E}^r \overline{\mathcal{Y}}_i^r(V_n^\infty) = 0, \tag{56b}$$

$$\lim_{n\to\infty} \mathbb{E}^r \overline{\mathcal{Y}}_i^r(H_n^\infty) = 0, \tag{56c}$$

where $V_n^\infty = [n,\infty)\times\mathbb{R}_+$ and $H_n^\infty = \mathbb{R}_+\times[n,\infty)$.

First check (56a). For each r, the route i population process $Z_i^r(\cdot)$ is bounded from above by the length $Q_i^r(\cdot)$ of the $M/G/\infty$ queue with the following parameters:

(Q.1) at $t=0$, there are $Z_i^r(0)$ customers whose service times are patience times of the initial flows on route i of the r-th network;

(Q.2) the input process is the route i input process of the r-th network;

(Q.3) service times of newly arriving customers are patience times of newly arriving flows on route i of the r-th network.

For all r and t, $Z_i^r(t) \le Q_i^r(t)$. As $t\to\infty$, $Z_i^r(t) \Rightarrow Y_i^r$ and $Q_i^r(t) \Rightarrow \Pi(\eta_i^r \mathbb{E}^r D_i^r)$. Hence, $Y_i^r \le_{\mathrm{st}} \Pi(\eta_i^r \mathbb{E}^r D_i^r)$ and $\mathbb{E}^r \overline{Y}_i^r \le \eta_i^r \mathbb{E}^r D_i^r/r \to \eta_i \mathbb{E} D_i$ as $r\to\infty$, which implies (56a).

Now check (56b). Note that, if at some point the residual flow size is at least n, then the initial flow size was at least n, too. Hence, $\mathcal{Z}_i^r(\cdot)(V_n^\infty)$ is bounded from above by the length $Q_i^{r,n}(\cdot)$ of the $M/G/\infty$ queue whose initial state is as in (Q.1), newly arriving customers are newly arriving flows on route i of the r-th network with initial sizes at least n, and service times of newly arriving customers are patience times of the corresponding flows. In particular, the input process for this queue is Poisson with intensity $\eta_i^r \mathbb{P}^r\{B_i^r \ge n\}$.

Let $f_n(\cdot)$ be a continuous function on $\mathbb{R}_+^2$ such that

$$\mathbb{I}_{V_{n+1}^\infty}(\cdot) \le f_n(\cdot) \le \mathbb{I}_{V_n^\infty}(\cdot).$$

Then, for all r and t,

$$\langle f_n, \mathcal{Z}_i^r(t)\rangle \le \mathcal{Z}_i^r(t)(V_n^\infty) \le Q_i^{r,n}(t)$$

Letting $t \to \infty$, we obtain

$$\mathcal{Y}_i^r(V_{n+1}^\infty) \le \langle f_n, \mathcal{Y}_i^r \rangle \le_{\rm st} \Pi(\eta_i^r \mathbb{P}^r\{B_i^r \ge n\}\mathbb{E}^r D_i^r),$$
$$\mathbb{E}^r \overline{\mathcal{Y}}_i^r(V_{n+1}^\infty) \le \eta_i^r \mathbb{P}^r\{B_i^r \ge n\}\mathbb{E}^r D_i^r/r,$$

and then (56b) follows.

Finally, (56c) is valid due to the following lemma.

Lemma 16. *For any $r \in \mathcal{R}$, i and Borel set $S \subseteq \mathbb{R}_+$,*

$$\mathcal{Y}_i^r(\mathbb{R}_+ \times S) \le_{st} \Pi(\eta_i^r\, \mathbb{E}^r D_i^r\, \mathbb{P}^r\{\tilde{D}_i^r \in S\}),$$

where $\tilde{D}_i^r$ has density $\mathbb{P}^r\{D_i^r > x\}/\mathbb{E}^r D_i^r$, $x \ge 0$.

Proof. Fix $r \in \mathcal{R}$, i and a Borel set $S \subseteq \mathbb{R}_+$. It suffices to show that, for any $\delta > 0$,

$$\mathcal{Y}_i^r(\mathbb{R}_+ \times S) \le_{\rm st} \Pi(\eta_i^r \mathbb{E}^r D_i^r \mathbb{P}^r\{\tilde{D}_i^r \in S^\delta\}),$$

so fix $\delta > 0$.

Consider the upper bound queue $Q_i^r(\cdot)$ with parameters (Q.1)–(Q.3). Denote by $Q_i^r(t)(S^\delta)$ the number of customers in this queue whose residual service times at time t are in S^δ. Then

$$\mathcal{Z}_i^r(\cdot)(\mathbb{R}_+ \times S^\delta) \le Q_i^r(\cdot)(S^\delta).$$

Given at time t there are k customers in the queue, denote by $D_1(t), \dots, D_k(t)$ their residual service times. By [26, Chapter 3.2, Theorem 2],

$$\lim_{t\to\infty} \mathbb{P}^r\{D_1(t) \le x_1, \dots, D_k(t) \le x_k | Q_i^r(t) = k\} = \mathbb{P}^r\{\tilde{D}_i^r \le x_1\} \dots \mathbb{P}^r\{\tilde{D}_i^r \le x_k\},$$

which together with $Q_i^r(t) \Rightarrow \Pi(\eta_i^r \mathbb{E}^r D_i^r)$ as $t \to \infty$ implies that

$$Q_i^r(t)(S^\delta) \Rightarrow \Pi(\eta_i^r \mathbb{E}^r D_i^r \mathbb{P}^r\{\tilde{D}_i^r \in S^\delta\}).$$

Let g_δ be a continuous function on $\mathbb{R}_+^2$ such that

$$\mathbb{I}_{\mathbb{R}_+ \times S}(\cdot) \le g_\delta(\cdot) \le \mathbb{I}_{\mathbb{R}_+ \times S^\delta}(\cdot).$$

Then, for any t,

$$\langle g^\delta, \mathcal{Z}_i^r(t) \rangle \le \mathcal{Z}_i^r(t)(\mathbb{R}_+ \times S^\delta) \le Q_i^r(t)(S^\delta),$$

and as $t \to \infty$,

$$\mathcal{Y}_i^r(\mathbb{R}_+ \times S) \le \langle g^\delta, \mathcal{Y}_i^r \rangle \le_{\rm st} \Pi(\eta_i^r \mathbb{E}^r D_i^r \mathbb{P}^r\{\tilde{D}_i^r \in S^\delta\}).$$ □

Appendix

Proof of Lemma 1. It suffices to show that, for a vector $z \in \mathbb{R}_+^I$ with the first $I' < I$ coordinates positive and the rest of them zero, and a sequence $\{z^k\}_{k\in\mathbb{N}} \subset (0,\infty)^I$ such that $z^k \to z$, we have $\Lambda(z^k) \to \Lambda(z)$.

Suppose that $z^k \to z$ but $\Lambda(z^k) \not\to \Lambda(z)$. Since $\{\Lambda(z^k)\}_{k\in\mathbb{N}}$ is a subset of the compact set $\{\Lambda \in \mathbb{R}_+^I \colon \|\Lambda\| \le \|C\|\}$, without loss of generality we may assume that $\Lambda(z^k) \to \Lambda' \ne \Lambda(z)$.

Recall that $\Lambda(z)$ is the unique optimal solution to

$$\text{maximize} \quad \sum\nolimits_{i=1}^I z_i \mathcal{U}_i(\Lambda_i/z_i) \quad \text{subject to} \quad A\Lambda \le C, \quad \Lambda \le m * z, \tag{57}$$

where, by convention, $\Lambda_i/0 := 0$ and $0 \times (-\infty) := 0$.

For all k, $A\Lambda(z^k) \le C$ and $\Lambda(z^k) \le m \cdot z^k$. Hence, Λ' is feasible for (57) and $\Lambda'_i = 0 = \Lambda_i(z)$ for $i > I'$. Since $\Lambda' \ne \Lambda(z)$ is not optimal for (57),

$$l := \sum\nolimits_{i=1}^{I'} z_i \mathcal{U}_i(\Lambda_i(z)/z_i) > \sum\nolimits_{i=1}^{I'} z_i \mathcal{U}_i(\Lambda'_i/z_i) =: r. \tag{58}$$

Now we construct a sequence $\Lambda^k \to \Lambda(z)$ such that Λ^k is feasible for the optimization problem (57) with z^k in place of z. Introduce vectors $C^k \in \mathbb{R}_+^J$ with $C_j^k = \sum_{i=I'+1}^{I} A_{ji}\Lambda_i(z^k)$. Put the first I' coordinates of Λ^k to be $\Lambda_i^k = (\Lambda_i(z) - \|C^k\|)^+ \wedge m_i z_i^k$, and the rest of them $\Lambda_i^k = \Lambda_i(z^k)$. That is, in the bandwidth allocation $\Lambda(z)$, the bandwidth C^k, which is required for the last $I - I'$ routes, is taken away from the first I' routes.

Since $z^k \to z$, $\Lambda^k \to \Lambda(z)$ and $\Lambda(z^k) \to \Lambda'$,

$$\sum\nolimits_{i=1}^{I'} z_i^k \mathcal{U}_i(\Lambda_i^k/z_i^k) \to l \quad \text{and} \quad \sum\nolimits_{i=1}^{I'} z_i^k \mathcal{U}_i(\Lambda_i(z^k)/z_i^k) \to r.$$

Also, for all k,

$$\sum\nolimits_{i=I'+1}^{I} z_i^k \mathcal{U}_i(\Lambda_i^k/z_i^k) = \sum\nolimits_{i=I'+1}^{I} z_i^k \mathcal{U}_i(\Lambda_i(z^k)/z_i^k).$$

Then, by (58), for k big enough,

$$\sum\nolimits_{i=1}^{I} z_i^k \mathcal{U}_i(\Lambda_i^k/z_i^k) > \sum\nolimits_{i=1}^{I} z_i^k \mathcal{U}_i(\Lambda_i(z^k)/z_i^k),$$

which contradicts to $\Lambda(z^k)$ being optimal for (57) with z^k in place of z. □

Proof of Corollary 1. Fix an FMS $(\zeta, z)(\cdot)$. By Theorem 3, $z(t) \to z^*$ as $t \to \infty$. Further we prove that $z(t) \to z^*$ implies $\zeta(t) \to \zeta^*$. It suffices to show that, for any $\varepsilon > 0$, there exists a t_ε such that, for all $t \ge t_\varepsilon$, i and Borel sets $A \subseteq \mathbb{R}_+^2$,

$$\begin{aligned} \zeta_i(t)(A) &\le \zeta_i^*(A^\varepsilon) + \varepsilon, \\ \zeta_i^*(A) &\le \zeta_i(t)(A^\varepsilon) + \varepsilon, \end{aligned} \tag{59}$$

so fix $\varepsilon > 0$.

For any $\delta \in (0, \min_{1\le i\le I} z_i^*)$, there exists a τ_δ such that, for all $t \ge \tau_\delta$,

$$z^* - \delta := (z_1^* - \delta, \dots, z_I^* - \delta) \le z(t) \le (z_1^* + \delta, \dots, z_I^* + \delta) =: z^* + \delta.$$

Then, for all $t \ge s \ge \tau_\delta$ and i, we have

$$\underbrace{r_i(z^* - \delta, z^* + \delta)}_{=: r_i^\delta}(t - s) \le S_i(z, s, t) \le \underbrace{R_i(z^* - \delta, z^* + \delta)}_{=: R_i^\delta}(t - s),$$

which, when plugged into the shifted fluid model equation (5a), implies that, for all $t \ge \tau_\delta$, i and Borel sets $A \subseteq \mathbb{R}_+^2$,

$$\zeta_i(t)(A) \le \zeta_i(\tau_\delta)(\mathbb{R}_+^2 \times [t - \tau_\delta, \infty)) + \eta_i \int_0^{t-\tau_\delta} \theta_i(A + (r_i^\delta s, s))ds, \tag{60a}$$

$$\zeta_i(t)(A) \ge \eta_i \int_0^{t-\tau_\delta} \theta_i(A + (R_i^\delta s, s))ds. \tag{60b}$$

Recall from Section 3 that, for all i and Borel sets $A \subseteq \mathbb{R}_+^2$,

$$\zeta_i^*(A) = \eta_i \int_0^{\infty} \theta_i(A + (\lambda_i(z^*)s, s))ds. \tag{61}$$

Now, there exists a t'_ε such that, for all i, Borel sets $A \subseteq \mathbb{R}^2_+$ and $\delta \in (0, \min_{1\le i\le I} z^*_i)$,

$$\eta_i \int_{t'_\varepsilon}^{\infty} \theta_i(A + (r^\delta_i s, s))ds \le \eta_i \int_{t'_\varepsilon}^{\infty} \mathbb{P}\{D_i \ge s\}ds \le \varepsilon/2, \tag{62a}$$

$$\eta_i \int_{t'_\varepsilon}^{\infty} \theta_i(A + (\lambda_i(z^*) s, s))ds \le \eta_i \int_{t'_\varepsilon}^{\infty} \mathbb{P}\{D_i \ge s\}ds \le \varepsilon/2. \tag{62b}$$

Take $\delta \in (0, \min_{1\le i\le I} z^*_i)$ such that

$$\|R^\delta - \lambda(z^*)\| t'_\varepsilon \le \varepsilon/2 \quad \text{and} \quad \|r^\delta - \lambda(z^*)\| t'_\varepsilon \le \varepsilon/2.$$

Then, for all i and Borel sets $A \subseteq \mathbb{R}^2_+$,

$$\eta_i \int_0^{t'_\varepsilon} \theta_i(A + (r^\delta_i s, s))ds \le \eta_i \int_0^{t'_\varepsilon} \theta_i(A^\varepsilon + (\lambda_i(z^*) s, s))ds \tag{63a}$$

$$\eta_i \int_0^{t'_\varepsilon} \theta_i(A + (\lambda_i(z^*) s, s))ds \le \eta_i \int_0^{t'_\varepsilon} \theta_i(A^\varepsilon + (R^\delta_i s, s))ds. \tag{63b}$$

Also take t''_ε such that, for all i,

$$\zeta_i(\tau_\delta)(\mathbb{R}^2_+ \times [t''_\varepsilon - \tau_\delta, \infty)) \le \varepsilon/2. \tag{64}$$

Now we put (60)–(64) together in order to obtain (59): for all $t \ge t_\varepsilon := (\tau_\delta + t'_\varepsilon) \vee t''_\varepsilon$, i and Borel sets $A \subseteq \mathbb{R}^2_+$,

$$\begin{aligned}\zeta_i(t)(A) &\overset{(60a),(64)}{\le} \varepsilon/2 + \eta_i \int_0^{t-\tau_\delta} \theta_i(A + (r^\delta_i s, s))ds \overset{(62a)}{\le} \varepsilon/2 + \eta_i \int_0^{t'_\varepsilon} \theta_i(A + (r^\delta_i s, s))ds + \varepsilon/2 \\ &\overset{(63a)}{\le} \eta_i \int_0^{t'_\varepsilon} \theta_i(A^\varepsilon + (\lambda_i(z^*) s, s))ds + \varepsilon \overset{(61)}{\le} \zeta^*_i(A^\varepsilon) + \varepsilon\end{aligned}$$

and

$$\begin{aligned}\zeta^*_i(A) &\overset{(61),(62a)}{\le} \eta_i \int_0^{t'_\varepsilon} \theta_i(A + (\lambda_i(z^*) s, s))ds + \varepsilon/2 \\ &\overset{(63b)}{\le} \eta_i \int_0^{t'_\varepsilon} \theta_i(A^\varepsilon + (R^\delta_i s, s))ds + \varepsilon/2 \overset{(60b)}{\le} \zeta_i(t)(A^\varepsilon) + \varepsilon.\end{aligned}$$

□

Proof of Lemma 4. For all $s \le t$ and $\varepsilon > 0$,

$$\begin{aligned}\int_s^t \mathbb{P}\{u + x \le \xi < u + x' + \varepsilon\}du &= \int_{s+x}^{t+x} \mathbb{P}\{\xi \ge u\}du - \int_{s+x'+\varepsilon}^{t+x'+\varepsilon} \mathbb{P}\{\xi \ge u\}du \\ &\le \int_{s+x}^{s+x'+\varepsilon} \mathbb{P}\{\xi \ge u\}du \le x' - x + \varepsilon.\end{aligned}$$

The lemma follows as we first let $\varepsilon \to 0$ (applying the dominated convergence theorem) and then $s \to -\infty$, $t \to \infty$. □

References

[1] Asmussen, S. *Applied probability and queues.* Springer, New York, 2003.

[2] Ayesta, U., Mandjes, M. Bandwidth-sharing networks under a diffusion scaling. *Annals of Operations Research*, 2009, Vol. 170, 41–58.

[3] Balder, E.J. Lecture notes on subdifferential calculus. `http://www.staff.science.uu.nl/~balde101/cao10/cursus10_1.pdf`

[4] Borst, S., Egorova, R., Zwart, B. Fluid limits for bandwidth-sharing networks in overload. Preprint, 2009.

[5] Bonald, T., Massoulié, L. Impact of fairness on Internet performance. *Proceedings of ACM Sigmetrics & Performance Conference*, 2001, Boston MA, USA, 82–91.

[6] Bonald, T., Massoulié, L., Proutière, A., Virtamo, J. A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing Systems*, 2006, Vol. 53, 65–84.

[7] Bramson, M. Stability of networks for max-min fair routing. Presentation at the *13th INFORMS Applied Probability Conference*, 2005, Ottawa ON, Canada.

[8] Chiang, M., Shan, D., Tang, A. Stochastic stability of network utility maximization: general file size distribution. *Proceedings of Annual Allerton Conference*, 2006, Monticello IL, USA.

[9] Egorova, R., Borst, S.C., Zwart, B. Bandwidth-sharing networks in overload. *Performance Evaluation*, 2007, Vol. 64, 978–993.

[10] Ethier, St. N., Kurtz, Th. G. *Markov processes: characterization and convergence.* John Wiley & Sons, New York, 1986.

[11] Gromoll, H.C., Robert, Ph., Zwart, B. Fluid limits for processor sharing queues with impatience. *Mathematics of Operations Research*, 2008, Vol. 33, 375–402.

[12] Gromoll, H.C., Williams, R.J. Fluid model for a data network with alpha-fair bandwidth sharing and general document size distributions: two examples of stability. *IMS Collections — Markov Processes and Related Topics*, 2008, Vol. 4, 253–265.

[13] Gromoll, H.C., Williams, R.J. Fluid limits for networks with bandwidth sharing and general document size distributions. *Annals of Applied Probability*, 2009, Vol. 19, 243–280.

[14] Jakubowski, A. Tightness criteria for random measures with application to the principle of conditioning in Hilbert spaces. *Probability and Mathematical Statistics*, 1988, Vol. 9.1, 95–114.

[15] Kallenberg, O. *Random Measures*. Akademie-Verlag, Berlin, 1983.

[16] Kang, W.N., Kelly, F.P., Lee, N.H., Williams, R.J. State space collapse and diffusion approximation for a network operating under a fair bandwidth-sharing policy. *Annals of Applied Probability*, 2009, Vol. 19, 1719–1780.

[17] Kang, W.N., Ramanan, K. Asymptotic approximations for stationary distributions of many-server queues with abandonment. *Annals of Applied Probability*, to appear.

[18] Kelly, F.P., Williams, R.J. Fluid model for a network operating under a fair bandwidth-sharing policy. *Annals of Applied Probability*, 2004, Vol. 14, 1055–1083.

[19] Kelly, F.P., Williams, R.J. Heavy traffic on a controlled motorway. *Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman*, 2010, 416–445.

[20] Massoulié, L. Structural properties of proportional fairness: stability and insensitivity. *Annals of Applied Probability*, 2007, Vol. 17, 809–839.

[21] Massoulié, L., Roberts, J.W. Bandwidth sharing: objectives & algorithms. *Proceedings of IEEE Infocom*, 1999, New York NY, USA, 1395–1403.

[22] Mo, J., Walrand, J. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 2000, Vol. 8, 556–567.

[23] Reed, J., Zwart, B. Limit theorems for bandwidth sharing networks with rate constraints. Submitted for publication, 2010.

[24] Roberts, J.W. A survey on statistical bandwidth sharing. *Computer Networks*, 2004, Vol. 45, 319–332.

[25] Roberts, J.W., Massoulié, L. Bandwidth sharing and admission control for elastic traffic. *Proceedings of ITC Specialist Seminar*, 1998, Yokohama, Japan.

[26] Taḱacs, L. *Introduction to the Theory of Queues*. Oxford University Press, New York, 1962.

[27] De Veciana, G., Lee, T.-L., Konstantopoulos, T. Stability and performance analysis of network supporting services with rate control — could the Internet be unstable? *Proceedings of IEEE Infocom*, 1999, New York NY, USA, 802–810.

[28] De Veciana, G., Lee, T.-L., Konstantopoulos, T. Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Transactions on Networking*, 2001, Vol. 9, 2–14.

[29] Ye, H., Yao, D.D. Heavy-traffic optimality of a stochastic network under utility-maximizing resource control. *Operations Research*, 2008, Vol. 56, 453–470.

[30] Ye, H., Yao, D.D. Utility maximizing resource control: diffusion limit and asymptotic optimality for a two-bottleneck model. *Operations Research*, 2010, Vol. 58, 613-623.